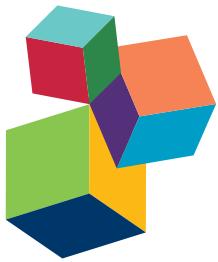


GAME CHANGER – NEXT GENERATION SEQUENCING AND ITS IMPACT ON FOOD MICROBIOLOGY

EDITED BY: Jennifer Ronholm, Sabah Bidawid and Sandra Torriani

PUBLISHED IN: *Frontiers in Microbiology*



frontiers

Frontiers Copyright Statement

© Copyright 2007-2018 Frontiers Media SA. All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714

ISBN 978-2-88945-463-1

DOI 10.3389/978-2-88945-463-1

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

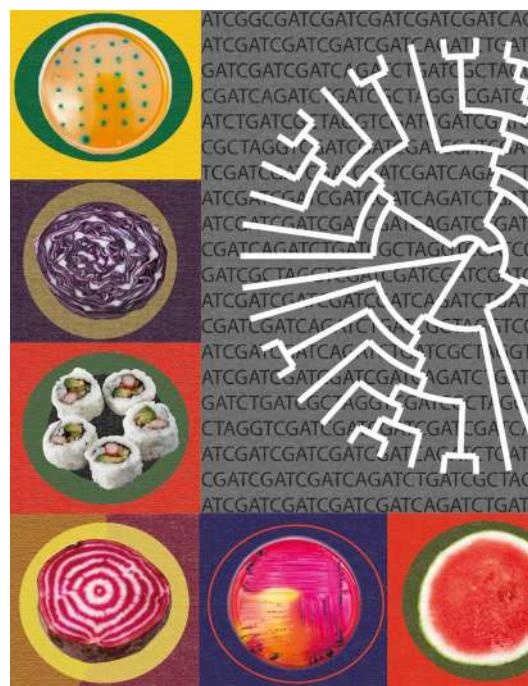
GAME CHANGER – NEXT GENERATION SEQUENCING AND ITS IMPACT ON FOOD MICROBIOLOGY

Topic Editors:

Jennifer Ronholm, McGill University, Canada

Sabah Bidawid, Health Canada, Canada

Sandra Torriani, University of Verona, Italy



The cover art was inspired by the work of Wassily Kandinsky. It is a stylistic take on food microbiology and the complex relationships between food, conventional culturing techniques, next generation sequencing, and phylogenetics.

Image created by Jennifer Ronholm, using photographs by Angela Catford.

Advances in next-generation sequencing technologies (NGS) are revolutionizing the field of food microbiology. Microbial whole genome sequencing (WGS) can provide identification, characterization, and subtyping of pathogens for epidemiological investigations at a level of precision previously not possible. This allows for connections and source attribution to be

inferred between related isolates that may be overlooked by traditional techniques. The archiving and global sharing of genome sequences allow for retrospective analysis of virulence genes, antimicrobial resistance markers, mobile genetic elements and other novel genes. The advent of high-throughput 16S rRNA amplicon sequencing, in combination with the advantages offered by massively parallel second-generation sequencing for metagenomics, enable intensive studies on the microbiomes of food products and the impact of foods on the human microbiome. These studies may one day lead to the development of reliable culture-independent methods for food monitoring and surveillance.

Similarly, RNA-seq has provided insights into the transcriptomes and hence the behaviour of bacterial pathogens in food, food processing environments, and in interaction with the host at a resolution previously not achieved through the use of microarrays and/or RT-PCR. The vast un-tapped potential applications of NGS along with its rapidly declining costs, give this technology the ability to contribute significantly to consumer protection, global trade facilitation, and increased food safety and security. Despite the rapid advances, challenges remain. How will NGS data be incorporated into our existing global food safety infrastructure? How will massive NGS data be stored and shared globally? What bioinformatics solutions will be used to analyse and optimise these large data sets?

This Research Topic discusses recent advances in the field of food microbiology made possible through the use of NGS.

Citation: Ronholm, J., Bidawid, S., Torriani, S., eds. (2018). Game Changer – Next Generation Sequencing and its Impact on Food Microbiology . Lausanne: Frontiers Media.
doi: 10.3389/978-2-88945-463-1

Table of Contents

- 07 Editorial: Game Changer – Next Generation Sequencing and Its Impact on Food Microbiology**
Jennifer Ronholm
- Chapter 1: Insights into Food Microbiology – Brought to you by Next Generation Sequencing**
- 10 Genomic Characterization of Dairy Associated Leuconostoc Species and Diversity of Leuconostocs in Undefined Mixed Mesophilic Starter Cultures**
Cyril A. Frantzen, Witold Kot, Thomas B. Pedersen, Ylva M. Ardö, Jeff R. Broadbent, Horst Neve, Lars H. Hansen, Fabio Dal Bello, Hilde M. Østlie, Hans P. Kleppen, Finn K. Vogensen and Helge Holo
- 24 Comparative Genomic Analysis Reveals Ecological Differentiation in the Genus Carnobacterium**
Christelle F. Iskandar, Frédéric Borges, Bernard Taminiau, Georges Daube, Monique Zagorec, Benoît Remenant, Jørgen J. Leisner, Martin A. Hansen, Søren J. Sørensen, Cécile Mangavel, Catherine Cailliez-Grimal and Anne-Marie Revol-Junelles
- 38 Genetic Characterization of the Exceptionally High Heat Resistance of the Non-toxic Surrogate Clostridium sporogenes PA 3679**
Robert R. Butler III, Kristin M. Schill, Yun Wang and Jean-François Pombert
- 49 Genotypes Associated with Listeria monocytogenes Isolates Displaying Impaired or Enhanced Tolerances to Cold, Salt, Acid, or Desiccation Stress**
Patricia Hingston, Jessica Chen, Bhavjinder K. Dhillon, Chad Laing, Claire Bertelli, Victor Gannon, Taurai Tasara, Kevin Allen, Fiona S. L. Brinkman, Lisbeth Truelstrup Hansen and Siyun Wang
- 69 A Genome-Wide Association Study to Identify Diagnostic Markers for Human Pathogenic Campylobacter jejuni Strains**
Cody J. Buchanan, Andrew L. Webb, Steven K. Mutschall, Peter Kruczakiewicz, Dillon O. R. Barker, Benjamin M. Hetman, Victor P. J. Gannon, D. Wade Abbott, James E. Thomas, G. Douglas Inglis and Eduardo N. Taboada
- 78 Evolution and Diversity of Listeria monocytogenes from Clinical and Food Samples in Shanghai, China**
Jianmin Zhang, Guojie Cao, Xuebin Xu, Marc Allard, Peng Li, Eric Brown, Xiaowei Yang, Haijian Pan and Jianghong Meng
- 87 Genome Sequence of Vibrio parahaemolyticus VP152 Strain Isolated from Penaeus indicus in Malaysia**
Vengadesh Letchumanan, Hooi-Leng Ser, Wen-Si Tan, Nurul-Syakima Ab Mutalib, Bey-Hing Goh, Kok-Gan Chan and Learn-Han Lee

Chapter 2: Whole Genome Sequencing and *Salmonella Enterica* – Novel Insights into an Important Pathogen

91 A Syst-OMICS Approach to Ensuring Food Safety and Reducing the Economic Burden of Salmonellosis

Jean-Guillaume Emond-Rheault, Julie Jeukens, Luca Freschi, Irena Kukavica-Ibrulj, Brian Boyle, Marie-Josée Dupont, Anna Colavecchio, Virginie Barrere, Brigitte Cadieux, Gitanjali Arya, Sadjia Bekal, Chrystal Berry, Elton Burnett, Camille Cavestri, Travis K. Chapin, Alanna Crouse, France Daigle, Michelle D. Danyluk, Pascal Delaquis, Ken Dewar, Florence Doualla-Bell, Ismail Fliss, Karen Fong, Eric Fournier, Eelco Franz, Rafael Garduno, Alexander Gill, Samantha Gruenheid, Linda Harris, Carol B. Huang, Hongsheng Huang, Roger Johnson, Yann Joly, Maud Kerhoas, Nguyet Kong, Gisèle Lapointe, Line Larivière, Stéphanie Loignon, Danielle Malo, Sylvain Moineau, Walid Mottawea, Kakali Mukhopadhyay, Céline Nadon, John Nash, Ida Ngueng Feze, Dele Ogunremi, Ann Perets, Ana V. Pilar, Aleisha R. Reimer, James Robertson, John Rohde, Kenneth E. Sanderson, Lingqiao Song, Roger Stephan, Sandeep Tamber, Paul Thomassin, Denise Tremblay, Valentine Usongo, Caroline Vincent, Siyun Wang, Joel T. Weadge, Martin Wiedmann, Lucas Wijnands, Emily D. Wilson, Thomas Wittum, Catherine Yoshida, Khadija Youfsi, Lei Zhu, Bart C. Weimer, Lawrence Goodridge and Roger C. Levesque

99 The Validation and Implications of Using Whole Genome Sequencing as a Replacement for Traditional Serotyping for a National *Salmonella* Reference Laboratory

Chris A. Yachison, Catherine Yoshida, James Robertson, John H. E. Nash, Peter Kruczakiewicz, Eduardo N. Taboada, Matthew Walker, Aleisha Reimer, Sara Christianson, Anil Nichani The PulseNet Canada Steering Committee and Celine Nadon

108 Pan-genome Analyses of the Species *Salmonella enterica*, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar

Chad R. Laing, Matthew D. Whiteside and Victor P. J. Gannon

124 Prophage Integrase Typing Is a Useful Indicator of Genomic Diversity in *Salmonella enterica*

Anna Colavecchio, Yasmin D'Souza, Elizabeth Tompkins, Julie Jeukens, Luca Freschi, Jean-Guillaume Emond-Rheault, Irena Kukavica-Ibrulj, Brian Boyle, Sadjia Bekal, Sandeep Tamber, Roger C. Levesque and Lawrence D. Goodridge

135 Temporal Genomic Phylogeny Reconstruction Indicates a Geospatial Transmission Path of *Salmonella Cerro* in the United States and a Clade-Specific Loss of Hydrogen Sulfide Production

Jasna Kovac, Kevin J. Cummings, Lorraine D. Rodriguez-Rivera, Laura M. Carroll, Anil Thachil and Martin Wiedmann

Chapter 3: Metagenomics and the Future of Next Generation Sequencing

147 Metagenomics: The Next Culture-Independent Game Changer

Jessica D. Forbes, Natalie C. Knox, Jennifer Ronholm, Franco Pagotto and Aleisha Reimer

168 Metagenomic Sequencing for Surveillance of Food- and Waterborne Viral Diseases

David F. Nieuwenhuijsen and Marion P. G. Koopmans

179 Characterization of the Genomic Diversity of Norovirus in Linked Patients Using a Metagenomic Deep Sequencing Approach

Neda Nasheri, Nicholas Petronella, Jennifer Ronholm, Sabah Bidawid and Nathalie Corneau

Chapter 4: First Insights into the Food Microbiome

- 193 *The Grapevine and Wine Microbiome: Insights from High-Throughput Amplicon Sequencing***
Horatio H. Morgan, Maret du Toit and Mathabatha E. Setati
- 208 *From Vineyard Soil to Wine Fermentation: Microbiome Approximations to Explain the “terroir” Concept***
Ignacio Belda, Iratxe Zarraonaindia, Matthew Perisin, Antonio Palacios and Alberto Acedo
- 220 *A Perspective Study of Koumiss Microbiome by Metagenomics Analysis Based on Single-Cell Amplification Technique***
Guoqiang Yao, Jie Yu, Qiangchuan Hou, Wenyang Hui, Wenjun Liu, Lai-Yu Kwok, Bilige Menghe, Tiansong Sun, Heping Zhang and Wenyi Zhang
- 231 *Genotyping by PCR and High-Throughput Sequencing of Commercial Probiotic Products Reveals Composition Biases***
Wesley Morovic, Ashley A. Hibberd, Bryan Zabel, Rodolphe Barrangou and Buffy Stahl
- 242 *Characterization of Gut Microbiome Dynamics in Developing Pekin Ducks and Impact of Management System***
Aaron A. Best, Amanda L. Porter, Susan M. Fraley and Gregory S. Fraley

Chapter 5: Where do we go from here?

- 257 *The Importance of Bacterial Culture to Food Microbiology in the Age of Genomics***
Alexander Gill
- 263 *Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance***
Emma Griffiths, Damion Dooley, Morag Graham, Gary Van Domselaar, Fiona S. L. Brinkman and William W. L. Hsiao
- 274 *Food Safety in the Age of Next Generation Sequencing, Bioinformatics, and Open Data Access***
Eduardo N. Taboada, Morag R. Graham, João A. Carriço and Gary Van Domselaar
- 284 *The Public Health Impact of a Publicly Available, Environmental Database of Microbial Genomes***
Eric L. Stevens, Ruth Timme, Eric W. Brown, Marc W. Allard, Errol Strain, Kelly Bunning and Steven Musser
- 288 *A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens***
Lee S. Katz, Taylor Griswold, Amanda J. Williams-Newkirk, Darlene Wagner, Aaron Petkau, Cameron Sieffert, Gary Van Domselaar, Xiangyu Deng and Heather A. Carleton



Editorial: Game Changer - Next Generation Sequencing and Its Impact on Food Microbiology

Jennifer Ronholm^{1,2*}

¹ Department of Animal Science, Faculty of Agricultural and Environmental Sciences, McGill University, Montreal, QC, Canada, ² Department of Food Science and Agricultural Chemistry, Faculty of Agricultural and Environmental Sciences, McGill University, Montreal, QC, Canada

Keywords: foodborne pathogens, food microbiology, whole genome sequencing, next generation sequencing, metagenomics

Editorial on the Research Topic

Game Changer - Next Generation Sequencing and Its Impact on Food Microbiology

In the decade between 2004 and 2014 the rapid evolution of next generation sequencing (NGS) platforms reduced the cost of sequencing a gigabase of nucleic acid from \$1,000 to \$10. This resulted in the widespread availability of DNA sequence data and started in a revolution in field of food microbiology.

A critical activity in maintaining microbiological food safety is fast and accurate outbreak detection and source attribution. Successfully delineating an outbreak relies primarily on a high-resolution comparison of the relatedness of clinical, food, and environmental samples. Traditional molecular techniques such as pulsed-field gel electrophoresis (PFGE) or multi-locus sequence typing (MLST) also rely on detecting subtle genomic differences between isolates, but neither technique is able to utilize the whole genome. Microbial whole genome sequencing (WGS) followed by detection of single nucleotide polymorphisms (SNPs) can extract relevant information from the entire genome and provides the highest-resolution examination of the relatedness of isolates possible. The genomic data that is generated in the process can also be mined for the presence of virulence factors, antibiotic resistance genes, or other genetic markers of interest. Therefore, WGS is rapidly displacing other molecular typing methods for foodborne outbreak analysis and surveillance.

Other sequencing techniques such as shot-gun metagenomics, single-cell sequencing, and 16S rRNA targeted-amplicon sequencing are routinely being used by microbial ecologists to characterize entire microbial communities. In a food microbiology context, these techniques are providing a detailed look at the microbiomes of food, food-animals, food-spoilage, fermentation, un-defined starter cultures, and probiotic products.

This e-book includes inputs from 242 authors who have used this Research Topic as a platform to discuss NGS as it relates to food microbiology including recent findings, novel applications, future directions, and concerns about the shift away from more traditional methods.

Twenty-two years ago, the first draft genome sequence of a bacteria (*Haemophilus influenzae*) was completed using a protocol that required individually Sanger sequencing the entire 1.8 million bp genome one 460 bp piece at a time. Each piece first had to be cloned into plasmid vectors and propagated in *Escherichia coli* cells (Fleischmann et al., 1995). This was a monumental task that required generating 9,500 *E. coli* clones to obtain a single draft genome with 5x coverage. Today, most well equipped microbiology laboratories can sequence, assemble, and annotate several bacterial draft genomes in a week.

OPEN ACCESS

Edited by:

Abd El-Latif Hesham,
Assiut University, Egypt

Reviewed by:

Rosanna Tofalo,
Università di Teramo, Italy

*Correspondence:

Jennifer Ronholm
jennifer.ronholm@mcgill.ca

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 20 November 2017

Accepted: 15 February 2018

Published: 09 March 2018

Citation:

Ronholm J (2018) Editorial: Game Changer - Next Generation Sequencing and Its Impact on Food Microbiology. *Front. Microbiol.* 9:363.
doi: 10.3389/fmicb.2018.00363

A whole genome sequence can provide identification, characterization, and subtyping of a microbe for epidemiological investigations at a speed and level of precision that was previously not possible (Ronholm et al., 2016). This precision allows regulatory agencies to make connections between clinical cases that may have been overlooked using traditional techniques. For example, *Salmonella enterica*, which is one of the most prevalent foodborne pathogens in the world (Scallan et al., 2011), is divided into >2,500 serovars, only a few of which are commonly associated with human illness. There is high genetic homogeneity between isolates belonging to the same serovar. Genetic homogeneity between serovars that regularly cause illness in humans creates a problem for public health organizations. For example, the majority of clinical *Salmonella* isolates received by the New York State Department of Health are *S. ser. Enteritidis*, and of these isolates, 50% are indistinguishable by pulsed field gel electrophoresis (PFGE) (Allard et al., 2013). The result is that from serotyping or PFGE data alone there is no way of knowing if there is a single *S. Enteritidis* outbreak occurring or if several simultaneous outbreaks are co-occurring—the resolution is simply not high enough to differentiate between outbreaks. WGS solves this problem and is therefore very useful, particularly in investigating *S. enterica* outbreaks. This is perhaps why four articles in research topic are dedicated to discussing *Salmonella* in the context of WGS. The Syst-OMICS consortium (<http://salmonella-sytemics.ca/en/>) is in the process of sequencing 4500 *Salmonella* genomes and is concurrently assessing the virulence potential of several of these isolates by cell-culture or *in vivo*. The Perspective article by Emond-Rheault et al. is the first publicly available report from this group. The intensive genomics analysis of the *Salmonella* genus presented by Emond-Rheault et al. is complemented, in this Research Topic, by the more focused article by Kovac et al. which addressed the genomics of a single serovar, *S. Cerro*. However, both articles reminded readers of the importance of correlating clinical outcome or wet-lab pathogenicity data with genomic data. Genetic mutations can lead to novel pheno- and patho-types, and studies that work through the process of gathering these complementary datasets provide invaluable information so that eventually accurate microbial risk assessments can be performed based only on genomic data.

Culture-independent diagnostic tests (CIDTs), which use nucleic acid and antigen-based assays, are beneficial to clinicians since they decrease diagnostic testing times and may even help in diagnosing cases that would have been overlooked (Huang et al., 2016). However, these techniques may be detrimental to public health, since without either an isolate or a whole genome sequence, the ability public health agencies to link clinical cases of a disease to each other and a source is severely limited. The review by Forbes et al. discussed metagenomics as a culture-independent solution to the public health short-falls of CIDTs in general, but reminded readers there are still challenges including generation non-target data, the inability to distinguish between live and dead bacterial cells, and the difficulty of assigning detected virulence and antibiotic resistance genes to a particular bacterial species. Single-cell metagenomics, still in the early stages of development,

may solve the latter problem. Indeed, in this Research Topic, Yao et al. used a single-cell amplification technique to successfully analyze the metagenome of koumiss, a traditional fermented dairy product.

The article by Nieuwenhuijse et al. suggested that since metagenomics can simultaneously detect all viruses from complex microbial sample this technique could be useful in viral surveillance. However, Nieuwenhuijse et al. also acknowledged that there are challenges with metagenomic viral detection including the excessive generation of non-target data and that the poor understanding of the virome makes data interpretation difficult. Despite these challenges, Nasheri et al. developed a technique that uses non-specific amplification of viral RNA, directly from patient fecal samples, to generate whole genome Norovirus sequences—a technique that could eventually be valuable for reconstructing transmission directionality.

NGS techniques have also allowed microbial ecologists to explore, compare, and characterize mixed microbial communities. Since ~98% of the bacteria in an environmental sample cannot be grown in the laboratory using routine techniques, this is really the first time these populations have been able to be effectively studied (Wade, 2002). The range of effects that a microbial community can have food-production, -safety, and -quality are essentially limitless. In this Research Topic four papers addressed a wide-range of microbiome related topics. NGS techniques have led to numerous recent studies into the effect of bacteria on the final sensory properties of wines. The review by Belda et al. introduced readers to influence of the soil microbiome on wine-quality, while Morgan et al. summarized the findings of several studies investigating the effects of the grape and vine microbiomes. The paper by Best et al. investigated the gastrointestinal microbiome of the Pekin Duck and while the data presented in this article provided only baseline information on the normal microbiome for this food animal, the author's also note that that there is potential for the microbiome of food animals to be optimized in the future to improve growth yield and exclude pathogens. The study by Morovic et al. used 16S rRNA sequencing to characterize a variety of probiotic supplements and discovered differences in both the classification and abundance of bacteria between the supplement and the bacteria named on the label. This article draws attention to the possible uses of NGS in detecting adulteration in probiotic supplements—which, given their findings, may be an important topic to address. While culture-independent sequencing techniques are critical for understanding the mixed microbial communities present in a food item, food-producing animal, or a probiotic supplement, the mini-review by Gill made the important point that the majority of microorganisms that are relevant to food-microbiology are easily cultured, and that genomics complements rather than replaces culture-based analysis when addressing most microbiological issues surrounding food production and distribution.

Despite the availability of NGS techniques and the amazing potential that sequencing data has in food safety management a few articles also reminded observant readers that there are still hurdles to be overcome before the full implementation of this technology by regulatory bodies. The perspective article

by Griffiths et al. introduced readers to the importance of metadata in understanding the context of whole genome sequences. Griffiths et al. informed readers that there are several important issues with metadata regarding language and consistency in major genomic databases and introduced us to ontologies—hierarchies of well-defined and standardized vocabularies interconnected by logical relationships, as a possible solution to this problem. To get the most from WGS for outbreak detection and surveillance global cooperation and data sharing is required. However, as noted in Taboada et al. several jurisdictions are hesitant about the rapid release of genomic data to public archives; and clarifying issues surrounding sensitivities of metadata, legal implications of increased source attribution accuracy, and intellectual property rights surrounding sequence data will be important in moving forward.

There were a few important topics related to both NGS and food microbiology that were not addressed in this issue. The most glaring omission is likely the exclusion of any mention of foodborne parasites. The further development of metagenomics has a huge potential for the detection, surveillance, and characterization of parasites, most of which are not culturable in the laboratory. In addition, a closed genome sequence has yet to be completed for most foodborne parasites, which is a goal several groups are actively pursuing. The Research Topic was limited to DNA sequencing and failed to address the

developing field of RNA-seq including its use as research tool, and its potential to help solve some of the existing issues with metagenomics for CIDT. Lastly, although researchers in the field of food microbiology openly discuss the difficulty they experience in recruiting highly qualified personnel in bioinformatics and microbial genomics and a few articles in this Research Topic alluded to this need, a study to systematically assess and quantify this need has yet to be performed and would be valuable to see in the future.

Overall the collection of articles in this Research Topic presents several of the benefits that embracing NGS would bring to the field of microbiology, while warning that, at least for now, genomic data complements rather than replaces *in vivo* virulence assays and culture-based studies of physiology. Several of issues surrounding data sharing, archiving, and analysis were also raised. I hope that research topic adequately informs readers about the benefits that NGS offers to the field of food microbiology and about the many challenges that have yet to be overcome in this field.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

REFERENCES

- Allard, M. W., Luo, Y., Strain, E., Pettengill, J., Timme, R., Wang, C., et al. (2013). On the evolutionary history, population genetics and diversity among isolates of *Salmonella enteritidis* PFGE pattern JEGX01.0004. *PLoS ONE* 8:e55254. doi: 10.1371/journal.pone.0055254
- Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512. doi: 10.1126/science.7542800
- Huang, J. Y., Henao, O. L., Griffin, P. M., Vugia, D. J., Cronquist, A. B., Hurd, S., et al. (2016). Infection with pathogens transmitted commonly through food and the effect of increasing use of culture-independent diagnostic tests on surveillance — foodborne diseases active surveillance network, 10 U.S. Sites, 2012–2015. *MMWR Morb. Mortal. Wkly. Rep.* 65, 368–371. doi: 10.15585/mmwr.mm6514a2
- Ronholm, J., Nasheri, N., Petronella, N., and Pagotto, F. (2016). Navigating microbiological food safety in the era of whole-genome sequencing. *Clin. Microbiol. Rev.* 29, 837–857. doi: 10.1128/CMR.00056-16

- Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., et al. (2011). Foodborne illness acquired in the United States—Major Pathogens. *Emerging Infect. Dis.* 17, 7–15. doi: 10.3201/eid1701. P11101
- Wade, W. (2002). Unculturable bacteria—the uncharacterized organisms that cause oral infections. *J. R. Soc. Med.* 95, 81–83. doi: 10.1258/jrsm.95.2.81

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Ronholm. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genomic Characterization of Dairy Associated *Leuconostoc* Species and Diversity of *Leuconostocs* in Undefined Mixed Mesophilic Starter Cultures

Cyril A. Frantzen¹, Witold Kot², Thomas B. Pedersen³, Ylva M. Ardö³, Jeff R. Broadbent⁴, Horst Neve⁵, Lars H. Hansen², Fabio Dal Bello⁶, Hilde M. Østlie¹, Hans P. Kleppen^{1,7}, Finn K. Vogensen³ and Helge Holo^{1,8*}

¹ Laboratory of Microbial Gene Technology and Food Microbiology, Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway, ² Department of Environmental Science, Aarhus University, Roskilde, Denmark, ³ Department of Food Science, University of Copenhagen, Copenhagen, Denmark, ⁴ Department of Nutrition, Dietetics and Food Sciences, Utah State University, Logan, UT, USA, ⁵ Department of Microbiology and Biotechnology, Max Rubner-Institut, Kiel, Germany, ⁶ Sacco Srl, Cordonago, Italy, ⁷ ACD Pharmaceuticals AS, Leknes, Norway, ⁸ TINE SA, Oslo, Norway

OPEN ACCESS

Edited by:

Sandra Torriani,
University of Verona, Italy

Reviewed by:

Beatriz Martínez,
Consejo Superior de Investigaciones
Científicas (CSIC), Spain
Daniel M. Linares,
Teagasc-The Irish Agriculture and
Food Development Authority, Ireland

*Correspondence:

Helge Holo
helge.holo@nmbu.no

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 17 November 2016

Accepted: 18 January 2017

Published: 03 February 2017

Citation:

Frantzen CA, Kot W, Pedersen TB,
Ardö YM, Broadbent JR, Neve H,
Hansen LH, Dal Bello F, Østlie HM,
Kleppen HP, Vogensen FK and Holo H
(2017) Genomic Characterization of
Dairy Associated *Leuconostoc*
Species and Diversity of
Leuconostocs in Undefined Mixed
Mesophilic Starter Cultures.
Front. Microbiol. 8:132.
doi: 10.3389/fmicb.2017.00132

Undefined mesophilic mixed (DL-type) starter cultures are composed of predominantly *Lactococcus lactis* subspecies and 1–10% *Leuconostoc* spp. The composition of the *Leuconostoc* population in the starter culture ultimately affects the characteristics and the quality of the final product. The scientific basis for the taxonomy of dairy relevant leuconostocs can be traced back 50 years, and no documentation on the genomic diversity of leuconostocs in starter cultures exists. We present data on the *Leuconostoc* population in five DL-type starter cultures commonly used by the dairy industry. The analyses were performed using traditional cultivation methods, and further augmented by next-generation DNA sequencing methods. Bacterial counts for starter cultures cultivated on two different media, MRS and MPCA, revealed large differences in the relative abundance of leuconostocs. Most of the leuconostocs in two of the starter cultures were unable to grow on MRS, emphasizing the limitations of culture-based methods and the importance of careful media selection or use of culture independent methods. Pan-genomic analysis of 59 *Leuconostoc* genomes enabled differentiation into twelve robust lineages. The genomic analyses show that the dairy-associated leuconostocs are highly adapted to their environment, characterized by the acquisition of genotype traits, such as the ability to metabolize citrate. In particular, *Leuconostoc mesenteroides* subsp. *cremoris* display telltale signs of a degenerative evolution, likely resulting from a long period of growth in milk in association with lactococci. Great differences in the metabolic potential between *Leuconostoc* species and subspecies were revealed. Using targeted amplicon sequencing, the composition of the *Leuconostoc* population in the five commercial starter cultures was shown to be significantly different. Three of the cultures were dominated by *Ln. mesenteroides* subspecies *cremoris*. *Leuconostoc pseudomesenteroides* dominated in two of the

cultures while *Leuconostoc lactis*, reported to be a major constituent in fermented dairy products, was only present in low amounts in one of the cultures. This is the first in-depth study of *Leuconostoc* genomics and diversity in dairy starter cultures. The results and the techniques presented may be of great value for the dairy industry.

Keywords: dairy, cheese, *leuconostoc*, comparative, genomics, diversity analysis, starter cultures, differentiation

INTRODUCTION

Mesophilic mixed (DL-type) starter cultures used in the production of Dutch-type cheeses are composed of undefined mixtures of homofermentative *Lactococcus lactis* subsp. *lactis* (*Lc. lactis*), *Lactococcus lactis* subsp. *cremoris* (*Lc. cremoris*), *Lactococcus lactis* subsp. *lactis* biovar. *diacetylactis* (*Lc. diacetylactis*) and heterofermentative *Leuconostoc* spp. The latter two provide aroma and texture by metabolizing citrate, producing diacetyl, acetoin and CO₂, while *Lc. cremoris* and *Lc. lactis* are the major acid producers through fermentation of lactose. In many cheeses, diacetyl is an important aroma compound, and CO₂ is important for the eye formation (Hugenholtz, 1993). In fermented dairy products, *Leuconostoc* grows in association with the acid-producing lactococci and have been suggested to play a role in promoting the growth of citrate positive *Lactococcus* strains (Vedamuthu, 1994; Bandell et al., 1998; Hache et al., 1999). The importance of *Leuconostoc* in cheese production is widely recognized. DL-type starter cultures are predominantly *Lactococcus* spp., *Leuconostoc* spp. commonly accounting for 1–10% of the starter culture population (Cogan and Jordan, 1994). However, knowledge on the species diversity of *Leuconostoc* included in these starter cultures, or the composition of *Leuconostoc* through the culture production is sparse. Due to the low initial number and relatively weak ability to ferment lactose, *Leuconostoc* spp. are not believed to have a significant effect in the acidification process in the early stages of cheese making (Ardö and Varming, 2010). However, leuconostocs have been shown to dominate the cheese microbiota in the later stages of ripening with added propionic acid bacteria (Porcellato et al., 2013; Østlie et al., 2016). The genus *Leuconostoc* is comprised of 13 species, with the species *Leuconostoc mesenteroides* divided into subspecies *mesenteroides*, *dextranicum*, *cremoris*, and *suionicum* (Hemme and Foucaud-Scheunemann, 2004; Gu et al., 2012). The *Leuconostoc* species (or subspecies) relevant for dairy production are *Leuconostoc mesenteroides* subsp. *mesenteroides* (*Ln. mesenteroides*), *Leuconostoc mesenteroides* subsp. *dextranicum* (*Ln. dextranicum*), *Leuconostoc mesenteroides* subsp. *cremoris* (*Ln. cremoris*), *Leuconostoc pseudomesenteroides* (*Ln. pseudomesenteroides*) and *Leuconostoc lactis* (*Ln. lactis*) (Cogan and Jordan, 1994; Thunell, 1995).

The bases for *Leuconostoc* taxonomy are results from cultivation-dependent methods, followed by phenotypic/biochemical characterization or non-specific molecular methods. In addition to being tedious and time-consuming, classical cultivation-dependent methods are known to underestimate the number of *Leuconostoc* spp., especially

Ln. cremoris (Vogensen et al., 1987; Ward et al., 1990; Auty et al., 2001). In addition, concerns on the lack of stability and reproducibility of phenotypical methods have been raised (Thunell, 1995; Barrangou et al., 2002). Several molecular typing methods, such as RAPD, PFGE, RFLP, Rep-PCR, MLST, MALDI-TOF MS, plasmid profiling and 16S rRNA targeted differentiation have been employed to characterize or identify *Leuconostoc* isolates (Villani et al., 1997; Björkroth et al., 2000; Cibik et al., 2000; Pérez et al., 2002; Sánchez et al., 2005; Viavainen and Björkroth, 2009; Nieto-Arribas et al., 2010; Alegria et al., 2013; Zeller-Péronnet et al., 2013; Dan et al., 2014; Zhang et al., 2015). However, most of these techniques requiring a preliminary stage of cultivation and comparison of results between the methods and between different laboratories remains challenging. Often, these methods were developed to work with only one or two species of *Leuconostoc*, so they do not provide subspecies differentiation, yield inconclusive results, yield results that are hard to reproduce, or provide arbitrary differentiation of isolates not sufficiently tethered to phenotypic traits. So far, the work by Dr. Ellen Garvie on the growth and metabolism of *Leuconostoc* spp. (Garvie, 1960, 1967, 1969, 1979, 1983; Garvie et al., 1974), and DNA-DNA hybridization studies (Farrow et al., 1989) remains the basis for the taxonomical division of dairy relevant leuconostocs.

The *Leuconostoc* genus has also not been subject to extensive genomic research, and information on the genomic diversity or species population dynamics through the cheese production processes is scarce if available at all. Scientific literature and product information on starter cultures pre-dating the genomic age list *Ln. cremoris* and *Ln. lactis* as the key *Leuconostoc* in undefined mixed mesophilic starter cultures (Lodics and Steenson, 1990; Johansen and Kibenich, 1992; Vedamuthu, 1994). However, in recent years, isolation of *Ln. mesenteroides*, *Ln. dextranicum*, and *Ln. pseudomesenteroides* is more common from starter cultures or from cheese derivatives (Olsen et al., 2007; Kleppen et al., 2012; Pedersen et al., 2014a,b; Østlie et al., 2016).

Here we present genomic comparative analysis of *Leuconostoc* spp. and present data on the diversity and composition of *Leuconostoc* populations in five commercially available DL-type starter cultures. Using traditional cultivation methods in combination with high-throughput sequencing techniques, we provide robust species and subspecies differentiation, and direct population composition analysis using targeted amplicon sequencing techniques. To our knowledge, this is the first in-depth genomic work performed on the *Leuconostoc* genus, and the first data published on *Leuconostoc* diversity in DL-type starter cultures.

MATERIALS AND METHODS

Cultivation of Bacterial Strains and Starter Cultures

All bacterial strains used in this study are listed in Supplementary Table S1. The two different media used for cultivation were de Man Rogosa Sharpe (MRS) (Difco, Detroit, Michigan, USA), and modified PCA (MPCA). PCA (Sigma-Aldrich, Oslo, Norway) was supplemented with 0.5 g/L Tween 80, 5.0 g/L ammonium-citrate, 1 g/L skim milk powder (TINE SA, Oslo, Norway), 0.04 g/L FeSO₄, 0.2 g/L MgSO₄, 0.05 g/L MnSO₄, and 10.0 g/L glucose. Glucose was sterile filtered separately and added after autoclaving. Both media were supplemented with 40 µg/mL vancomycin to select for *Leuconostoc*. Three separate extractions from one batch of each starter cultures (A, B, C, D, and E) were suspended in MPCA to an optical density at 600 nm (OD₆₀₀) of 1.0, serially diluted in 10% (w/v) skim milk and spread plated on MRS and MPCA agar plates in triplicate. The plates were incubated at 22°C for 5 days before colony enumeration. Isolates were transferred to MRS and MPCA broth media, respectively, and cultivated at 22°C for two passages before aliquots were supplemented with 15% (w/v) glycerol (Sigma-Aldrich) and stored at -70°C.

Genome Sequencing, Assembly, and Annotation

Genomic DNA from *Leuconostoc* isolates was extracted from 1 mL of overnight culture using Qiagen DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany). The cells were lysed with 40 mg/mL lysozyme (Qiagen, Hilden, Germany) and bead-beating in a FastPrep®-24 (MP Biomedicals, Santa Ana, California) using 0.5 g acid-washed beads (<106 µm) (Sigma-Aldrich) prior to column purification. DNA libraries were made using the Nextera XT DNA Sample Prep kit (Illumina, San Diego, California, USA) according to manufacturer instructions and sequenced with Illumina MiSeq (Illumina, San Diego, California, USA) using V3 chemistry for 33 isolates sequenced at the Norwegian University of Life Sciences, and V2 chemistry for 13 isolates sequenced at the Aarhus University. Raw sequences were adapter trimmed, quality filtered (Q>20), *de novo* assembled using SPAdes V3.7.1 (Nurk et al., 2013) and annotated using the Prokka pipeline (Seemann, 2014). Contigs shorter than 1000 bp or with < 5 times coverage were removed from each assembly prior to gene annotation. Thirteen publicly available genomes of *Leuconostoc* obtained from the National Center for Biotechnology Information (NCBI) database were also included in the dataset (Jung et al., 2012; Meslier et al., 2012; Erkus et al., 2013; Pedersen et al., 2014a,b; Campedelli et al., 2015; Østlie et al., 2016). This whole genome project has been deposited at DDBJ/ENA/GenBank under the BioProject PRJNA352459.

Genomic Analysis

The protein coding sequences of all *Leuconostoc* isolates were compared by an all-against-all approach using BLASTP (Camacho et al., 2009) and grouped into orthologous clusters using GET_HOMOLOGUES (Version 2.0.10) (Contreras-Moreira and Vinuesa, 2013). Pan and core genomes were estimated using the pan-genomic analysis tool PanGP

v.1.0.1 (Zhao et al., 2014). Orthologous groups (OGs) were identified via the Markov Cluster Algorithm (MCL) with an inflation value of 1.5 (Enright et al., 2002) and intersected using the compare_clusters.pl script provided with GET_HOMOLOGUES. The orthologous clusters were curated to exclude significantly divergent singletons, which is likely the result of erroneous assembly or annotation. A presence/absence matrix for each gene cluster and each genome was constructed for the pan-genome before statistical and clustering analysis of the matrix was performed in R (<http://www.r-project.org/>). Hierarchical clustering of the pan-genome matrix was performed using complete-linkage UPGMA with Manhattan distances, and a distance cut-off for the number of clusters was determined using the knee of the curve approach (Salvador and Chan, 2004), binning the isolates into genomic lineages. The resulting distance-matrix was used to construct a heatmap with dendograms using the heatmap.2 function included in the Gplots package (Version 2.16; Warnes et al., 2015) supplemented by the Dendextend package (Version 0.18.3; Galili, 2015).

Comparative Genomics Analysis

The genetic potential of individual *Leuconostoc* lineages that were identified by the pan-/core-genome analysis was investigated by producing intra-lineage pan-genomes using GET_HOMOLOGUES (Version 2.0.10). The pan-genome for each lineage was analyzed using Blast2GO v4 (Conesa et al., 2005) to identify functionality, and Geneious 8.1.8 (Kearse et al., 2012) to identify sequence variation within orthologous clusters. The lineage pan-genomes were then compared using KEGG databases (Kanehisa and Goto, 2000) and the functional comparative comparison tool found in The SEED Viewer (Overbeek et al., 2014). CRISPR sequences and spacers were identified using the CRISPRFinder tool (Grissa et al., 2007).

Relative Quantification of *Leuconostoc* Species in Starter Cultures

Compositional analysis of *Leuconostoc* in five commercially available starter cultures was performed in triplicates on total DNA isolated from the starter cultures using 1 mL of starter culture diluted to an OD₆₀₀ of 1. The cultures were treated with 20 mg/mL lysozyme (Sigma-Aldrich) and 3U/L mutanolysin (Sigma-Aldrich), mechanically lysed using FastPrep (MP Biomedicals) with 0.5 g of acid-washed beads (<106 µm) (Sigma-Aldrich) and purified using the Qiagen DNeasy Blood & Tissue Kit (Qiagen). A suitable amplicon target was identified by screening the core-genome for nucleotide sequence variation using the sequence alignment metrics functions available in the DECIPIER package v1.16.1 (Wright, 2015). Genes without flanking consensus regions within a 500 bp variable region adequate for differentiation, or did not provide sufficient discrimination from similar sequences in species likely to be present in dairy, were excluded. The locus *eno* encoding for enolase was amplified by PCR using the KAPA HiFi PCR Kit (KAPA Biosystems, Wilmington, Massachusetts, USA) with primers Eno-F (5'-AACACGAAGCTGTGAATTGCGTG-3'), and Eno-R (5'-GCAAATCCACCTTCATCACCAACTGA-3'). Forward (5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-) and

reverse (5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGA CAG-) Illumina adapter overhangs were added to the 5' end of the primers to allow for Nextera XT DNA indexing of the PCR-products. The resulting libraries were sequenced on an Illumina MiSeq with V3 (2×300 bp) reagents. The resulting data were paired-end-joined and quality filtered using PEAR (Zhang et al., 2014) and clustered with a 100% identity level threshold using usearch v7 (Edgar, 2010) with error-minimization from uparse (Edgar, 2013). The resulting sequences were matched against a local BLAST-database produced from the *Leuconostoc* genomes for identification.

RESULTS

Leuconostoc in Dairy Starters

Enumeration on MRS-agar has been reported to underestimate the number of leuconostocs, especially *Ln. cremoris* (Vogensen et al., 1987; Ward et al., 1990; Auty et al., 2001). Bacterial counts were compared in five starter cultures (A, B, C, D, and E) commonly used in the production of Dutch-type cheeses using MRS and MPCA agar with 40 $\mu\text{g}/\text{mL}$ vancomycin. The results (Figure 1) showed large differences in the counts between starter cultures for the two media. Cultures A and D gave substantially higher counts on MPCA compared to MRS, while cultures B, C, and E had similar counts on both media. Thus, cultures A and D seemed to contain a large number of *Leuconostoc* strains unable to grow on MRS, while cultures B, C, and E did not.

Genome Sequencing and Pan-Genomic Analysis

Leuconostoc diversity was investigated by whole-genome sequencing of 20 isolates picked from MPCA- and MRS-plates

of cultures A and D, and 26 isolates from cheese, including Dutch-type cheese produced using cultures B, C, and E. Lastly, 13 publicly available *Leuconostoc* spp. genomes were included in the dataset. All 59 *Leuconostoc* genomes were annotated and the coding sequences (CDS) were compared by a blast-all-against-all approach to identify OGs. Pan- and core-genomes were estimated (Figure 2) using the pan-genomic analysis tool PanGP. After curation, the pan-genome was determined to consist of 4415 OGs, and a core-genome was found to comprise 638 OGs. Differentiation of isolates using hierarchical clustering on the pan-matrix clearly separated *Leuconostoc* species and sub-species (Figure 3). Several of the strains previously identified as *Ln. mesenteroides* subspecies were shown to be *Ln. pseudomesenteroides* by the genomic analysis. Moreover, the NCBI strain LbT16 previously identified as *Ln. cremoris*, was an outlier to the *Ln. cremoris* species branch and was identified in the pan-genomic analysis as *Ln. mesenteroides*. This was further confirmed by alignment of the full-length 16S rRNA, revealing a 100% identity between *Ln. cremoris* LbT16 and *Ln. mesenteroides* type 16S rRNA. Based on sequence similarity and gene content, the pan-genomic clustering divided the 59 leuconostocs into 12 robust *Leuconostoc* lineages across the genus. These included three lineages of *Ln. cremoris* (C1-C3), four lineages of *Ln. pseudomesenteroides* (P1-P4), four lineages of *Ln. mesenteroides* (M1-M4), and one lineage of *Ln. lactis* (L1). The *Ln. cremoris* TIFN8 genome was excluded from further analysis because the genome data contained a high number of fragmented genes and redundant sequences, making it an outlier.

The differences between lineages (Table 1), species and subspecies level (in the case for *Ln. mesenteroides* subsp.) include significantly smaller genomes for *Ln. cremoris* and *Ln. lactis* (1.6–1.8 Mb) compared to *Ln. mesenteroides*, *Ln. dextranicum*,

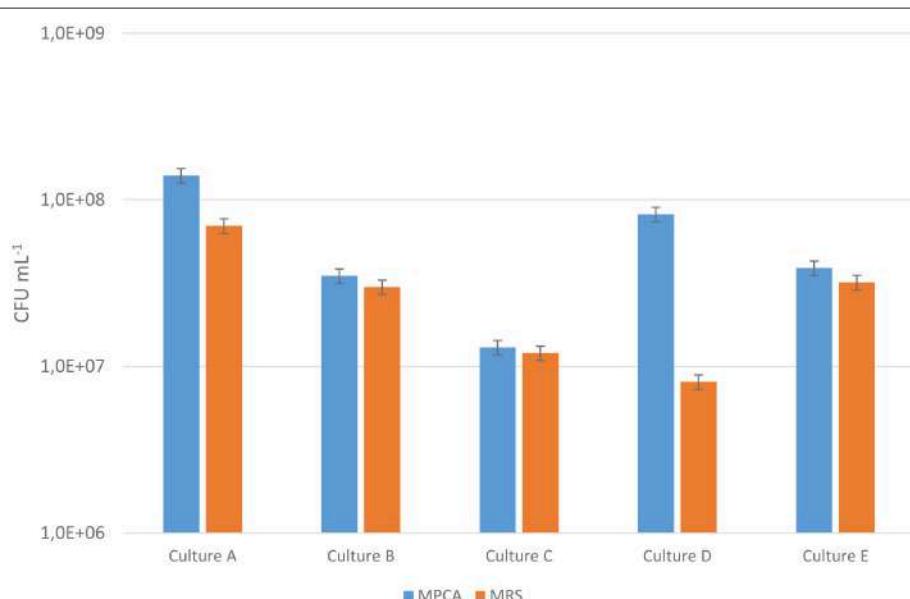
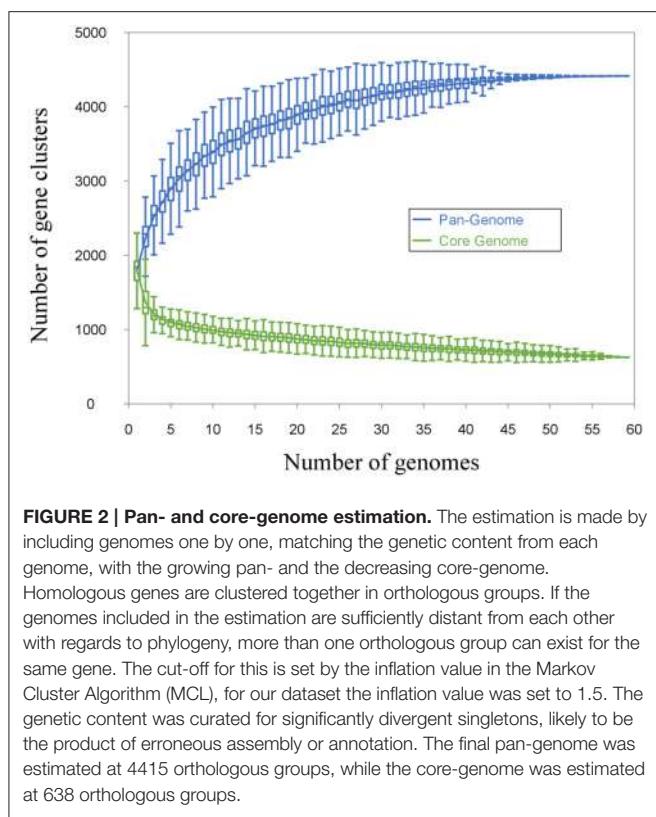


FIGURE 1 | Bacterial counts for five starter cultures A–E on MRS and MPCA supplemented with vancomycin to select for *Leuconostoc*. The counts are the mean of three separate extractions made from the same culture batch and the error bar indicates the standard deviation. The blue bars represent the bacterial counts on MPCA, while the orange bars represent the bacterial counts on MRS. The Y-axis is cut at 1,0E+06 for better readability.



and *Ln. pseudomesenteroides* (1.8–2.2 Mb). Moreover, the larger genome found in the latter three species contained up to 400 more coding sequences (CDS) than *Ln. cremoris* and *Ln. lactis*. Analysis of functional genomics indicated a closer relationship between *Ln. lactis* and *Ln. pseudomesenteroides*, than that of *Ln. mesenteroides*. Comparison of genetic potential within and between the *Ln. mesenteroides* subspecies showed only minor differences between *Ln. mesenteroides* and *Ln. dextranicum*. Rather, as shown in Figure 3, the variation between the isolates was much greater than the difference between *Ln. mesenteroides* and *Ln. dextranicum*. On the other hand, substantial difference was found between isolates of dairy origin and non-dairy origin. This environment adaptation was also observed for *Ln. lactis*, where *Ln. lactis* 91922, isolated from kimchi was clearly distinguishable from LN19 and LN24 isolated from dairy. Comparison of *Ln. cremoris* and other *Ln. mesenteroides* subspecies isolates revealed that a range of genetic elements found in these species that were missing in *Ln. cremoris*. Apart from some enzymes encoding for rhamnose-containing glucans, *Ln. cremoris* isolates did not have any genetic functionality absent in *Ln. mesenteroides* or *Ln. dextranicum*. Moreover, several truncated genes and deletions were found in *Ln. cremoris* isolates, likely the result of a degenerative evolutionary process through a long period of growth in the milk environment.

Comparative Genomics of Intra-Species *Leuconostoc* Lineages

To explore differences in functional genetic potential between the lineages within the species and subspecies, comparative analysis

of intra-lineage pan-genomes was performed. The results are included in Supplementary Table S2.

(I) *Ln. cremoris* Lineages

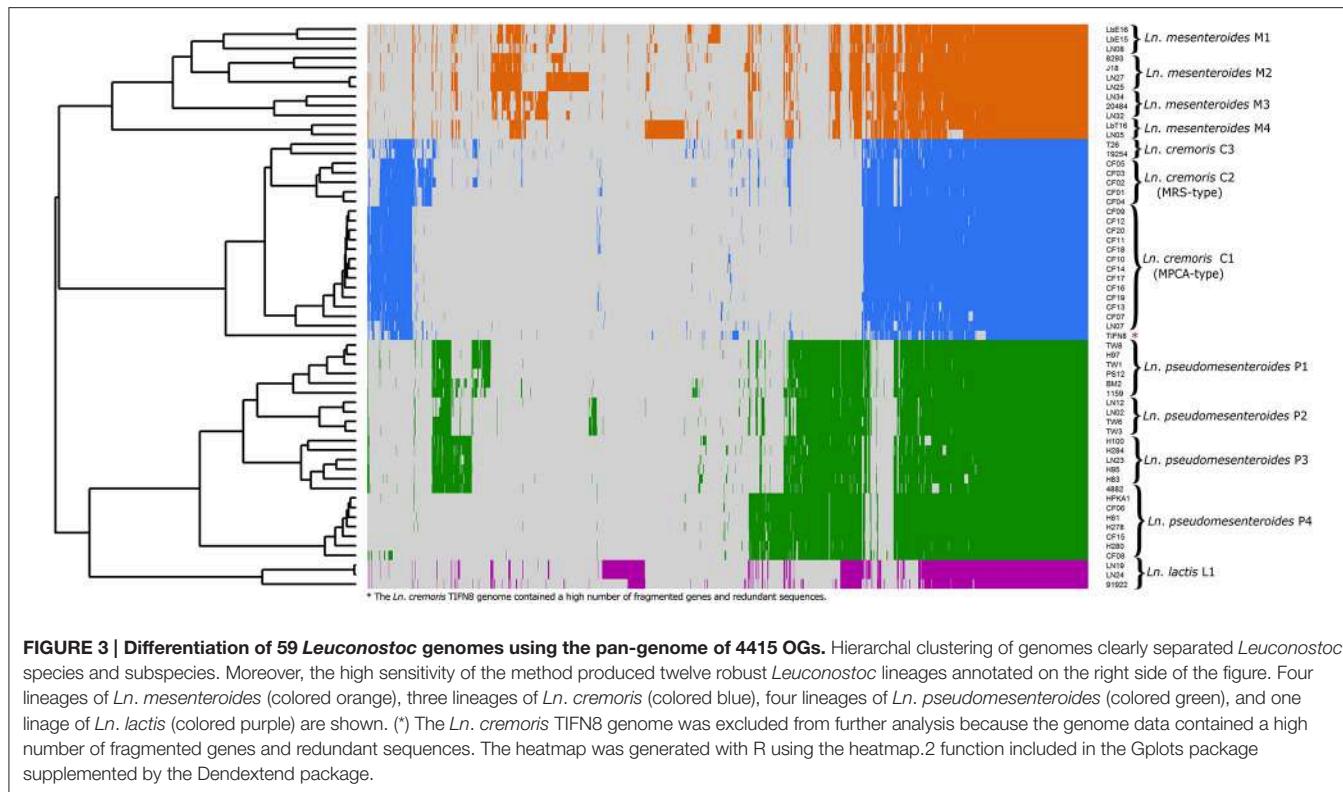
Comparison of the genetic content for *Ln. cremoris* lineages showed that *Ln. cremoris* C1, C2, and C3 were highly similar and differentiated from each other mostly because of sequence variation in shared OGs. *Ln. cremoris* C1 (MPCA-type), which did not grow on MRS was missing four OGs found in both lineage C2 and C3 (MRS-type). These OGs were annotated *rmlA*, *rmlB*, *rmlC*, and *rmlD*, encoding for four enzymes identified in the subsystem “rhamnose containing glycans.” These enzymes are associated with polysaccharide biosynthesis and their presence likely does not explain the inability of C1-type strains to grow on MRS.

(II) *Ln. mesenteroides* and *Ln. dextranicum* Lineages

Comparison of the genetic content showed a large variance between and within the *Ln. mesenteroides* lineages. Interestingly, no major difference between subspecies *Ln. mesenteroides* and *Ln. dextranicum* was found. *Ln. dextranicum* 20484 is grouped together with *Ln. mesenteroides* isolates LN32 and LN34, while *Ln. dextranicum* LbE16 is grouped together with *Ln. mesenteroides* LbE15 and LN08. This subspecies segregation of *Ln. dextranicum* and *Ln. mesenteroides* was based on the phenotypical ability to produce dextran from sucrose. Dextranucrase, the enzyme involved in this process, is a glucosyltransferase that catalyzes the transfer of glucosyl residues from sucrose to a dextran polymer and releases fructose. Several glucosyltransferases were found within all *Ln. mesenteroides* isolates included in this study, among them several genes encoding for dextranucrases with 40–67% amino acid identity to each other. Genotypically, the potential for dextran production exists within many if not all *Ln. mesenteroides* isolates, and does not differentiate *Ln. mesenteroides* from *Ln. dextranicum*. This finding was manifest by the separation of *Ln. mesenteroides* and *Ln. dextranicum* isolates into four lineages. Functional comparative analyses showed that the presence of the *cit* operon necessary for metabolism of citrate, and the *lacLM* genes is a characteristic of dairy-associated *Ln. mesenteroides*, *Ln. cremoris* and *Ln. pseudomesenteroides*. In all of the strains in lineages M3 and M4, both the *cit* operon and the *lacLM* genes were present, while strains in lineages M1 and M2 were lacking the *cit* operon, and half of them also lacked the *lacLM* genes. Furthermore, the strains in lineages M1 and M2 contained the genetic potential for metabolism of arabinose, and the two isolates J18 and ATCC8293 also contained genetic potential for xylose and β-glucoside metabolism. The lineage M4 strains LbT16 and LN05 also contained the deletion in the *lacZ* gene which is commonly identified in *Ln. cremoris* type strains. A genetic potential for proteolysis of casein (*prtP*) was identified in *Ln. mesenteroides* lineages M1 and M4, but not in M2 or M3.

(III) *Ln. lactis* Lineages

The pan-genomic differentiation grouped all the *Ln. lactis* isolates into one lineage. However, differences in genetic potential were found between the kimchi isolate *Ln. lactis* 91922 and



dairy isolates LN19 and LN24. *Ln. lactis* 91922 lacked citrate metabolism genes *citCDEFG*, but carried genetic potential for a maltose and glucose specific PTS system, metabolism of arabinose and a CRISPR-Cas operon, that were not found in the other two *Ln. lactis* isolates.

(IV) *Ln. pseudomesenteroides* Lineages

Despite the significant pan-genomic differences and the sequence variation in shared OGs, the functional differences between lineages of *Ln. pseudomesenteroides* were surprisingly few. *Ln. pseudomesenteroides* P4 was different from the other three lineages with regards to genome synteny and genetic potential. Genetic functionality in the category of methionine biosynthesis, β -glucoside metabolism, sucrose metabolism, as well as an additional lactate dehydrogenase was identified in *Ln. pseudomesenteroides* P4 but not P1, P2, and P3. Moreover, P4 isolates were missing the genes for reduction of diacetyl to acetoin and 2,3-butandiol, and contained genes for a different capsular and extracellular polysaccharide biosynthesis pathway, compared to P1, P2, and P3 isolates.

Genetic Potential of *Leuconostoc*

(I) Amino Acid Biosynthesis

The amino acid requirements of leuconostocs have been described as highly variable between strains. Glutamic acid and valine are required by most leuconostocs, methionine usually stimulates growth, while no *Leuconostoc* are reported to require alanine (Garvie, 1967). Comparative analysis of genes involved in amino acid biosynthesis showed that *Ln. cremoris*

and *Ln. mesenteroides* subspecies carried the genetic potential to produce a wide range of amino acids while *Ln. lactis* and *Ln. pseudomesenteroides* did not (Table 2). This included genes encoding biosynthesis of histidine, tryptophan, methionine and lysine. Studies on the amino acid requirement of leuconostocs show that most of the *Ln. mesenteroides* subspecies do require isoleucine and leucine to grow. The *ilv* and *leu* operons involved in biosynthesis of the branched-chain amino acids isoleucine, leucine and valine were present in all *Ln. mesenteroides* isolates, however both operons were truncated when compared to functional *ilv* and *leu* operons from lactococci. The *leuA* gene in the *leuABCD* operon is truncated in leuconostocs (391 aa) compared to lactococci (513 aa) likely resulting in an inactive product and a nonfunctional pathway. This has been documented in the dairy strain *Lactococcus lactis* IL1403 where a similar truncation of the *leuA* gene led to an inactivation of the leucine/valine pathway (Godon et al., 1993). Likewise, the *ilv* operon of sequenced leuconostocs is missing the *ilvD* gene, and has truncated *ilvA* and *ilvH* genes when compared to the lactococcal *ilv* operon. The truncation of *ilvA* has been shown to result in inactivation of the product, and would by itself be sufficient to abort the biosynthesis pathway (Cavin et al., 1999). None of the leuconostocs had genes for biosynthesis of glutamic acid. *Ln. lactis* isolates also lacked the genetic potential for cysteine biosynthesis.

(II) Carbohydrate Metabolism

Differences in the genetic potential within and between the *Leuconostoc* species were analyzed by comparing intra-species

TABLE 1 | Average genome size and coding sequences of *Leuconostoc* isolates binned into pan-genome lineages.

Profile name	Average genome size (Mb)	Average CDS
<i>Ln. cremoris</i> C1 (MPCA-type)	1.680 (± 5)	1760 (± 20)
<i>Ln. cremoris</i> C2 (MRS-type)	1.741 (± 40)	1822 (± 30)
<i>Ln. cremoris</i> C3	1.765 (± 124)	1956 (± 198)
<i>Ln. mesenteroides</i> M1	1.869 (± 19)	1851 (± 7)
<i>Ln. mesenteroides</i> M2	2.150 (± 123)	2212 (± 162)
<i>Ln. mesenteroides</i> M3	2.014 (± 19)	2074 (± 18)
<i>Ln. mesenteroides</i> M4	2.061 (± 219)	2101 (± 173)
<i>Ln. pseudomesenteroides</i> P1	2.028 (± 47)	2081 (± 61)
<i>Ln. pseudomesenteroides</i> P2	1.921 (± 25)	1925 (± 46)
<i>Ln. pseudomesenteroides</i> P3	2.063 (± 44)	2133 (± 60)
<i>Ln. pseudomesenteroides</i> P4	2.032 (± 61)	2046 (± 60)
<i>Ln. lactis</i> L1	1.718 (± 26)	1700 (± 43)

Information on each individual isolate is included in Supplementary Table S1.

pan-genomes using Blast2GO and the Seed Viewer. The *Leuconostoc* genus is composed of heterofermentative bacteria that use the phosphoketolase pathway to ferment hexoses. Therefore, it was not surprising to find that none of the isolates contained the gene for phosphofructokinase, a key enzyme in the Embden-Meyerhof pathway. However, a gene encoding fructose-bisphosphate aldolase class II was present in *Ln. lactis* and *Ln. pseudomesenteroides*. This could indicate a potential for synthesis of fructose-1,6-bisphosphate and glyceraldehyde-3-phosphate through fructose-1-phosphate, and hence homofermentative breakdown of fructose in *Ln. lactis* and *Ln. pseudomesenteroides*.

Comparative analysis of genes related to carbohydrate metabolism revealed big differences between the species (Table 3). All leuconostocs in this study encode beta-galactosidase, enabling utilization of lactose. Interestingly, the dairy *Ln. mesenteroides* have two different beta-galactosidases, *lacZ* and the plasmid-encoded *lacLM* (Obst et al., 1995), while the non-dairy isolates only contain *lacZ*. In *Ln. cremoris*, *lacZ* contains a large central deletion of 1200 bp between positions 740–1940. The *Ln. lactis* isolates only encode beta-galactosidase through *lacZ*, while the *Ln. pseudomesenteroides* isolates only encode beta-galactosidase through *lacLM*. In *Leuconostoc*, lactose is taken up by the lactose-specific transporter LacS, which couples lactose uptake to the secretion of galactose. LacS contains a C-terminal EIIAGlc-like domain and in *S. thermophilus* it has been shown that this domain can be phosphorylated, causing an increased lactose uptake rate (Gunnewijk and Poolman, 2000). All *Leuconostoc* isolates have this gene, but in *Ln. cremoris* *lacS* is truncated and lacks the C-terminal domain, possibly affecting lactose uptake and hence, growth rate on lactose. Alignment of all *lacS* sequences from this study revealed a close relationship between *Ln. pseudomesenteroides*, *Ln. lactis*, and *Ln. mesenteroides* isolates of non-dairy origin. In fact, *lacS* of non-dairy associated *Ln. mesenteroides* is more similar to the *lacS* from *Ln. lactis* and *Ln. pseudomesenteroides* ($>75\%$ identity) than that of dairy-associated *Ln. mesenteroides* or

TABLE 2 | Presence of genes encoding enzymes for amino acid biosynthesis.

Amino acid pathway	<i>Ln. cremoris</i>	<i>Ln. mesenteroides</i>	<i>Ln. lactis</i>	<i>Ln. pseudomesenteroides</i>
Alanine	+	+	+	+
Arginine	+	+	+	+
Aspartate	+	+	+	+
Cysteine	+	+	–	+
Glutamine	–	–	+	+
Glutamic acid	–	–	–	–
Glycine	+	+	+	+
Histidine	+	+	–	–
Isoleucine	–	–	–	–
Leucine	–	–	–	–
Lysine	+	+	+	–
Methionine	+	+	–	–
Phenylalanine	+	+	+	+
Proline	+	+	+	+
Serine	+	+	+	+
Threonine	+	+	+	+
Tryptophan	+	+	–	–
Tyrosine	+	+	+	+
Valine	–	–	–	–

+, presence of predicted pathway functionality; –, absence of predicted pathway functionality.

Ln. cremoris ($<36\%$ identity). Genes coding for maltose phosphorylase (*malP*) and sucrose-6-phosphate hydrolase (*scrB*) were found in *Ln. lactis*, *Ln. pseudomesenteroides* P4, and *Ln. mesenteroides*, but not *Ln. cremoris*. These enzymes are central to the metabolism of maltose and sucrose. Isolates containing *malP* also contained genes *malR* and *mall*, as well as a maltose epimerase. *Ln. lactis* and *Ln. pseudomesenteroides* also contained the *malEFG* gene cluster encoding for an ABC transporter, however the *malEFG* genes were truncated in *Ln. pseudomesenteroides*. Genes encoding for β -glucosidase (*bglA*) enabling utilization of salicin and arbutin was found in all *Ln. pseudomesenteroides* and *Ln. lactis* isolates, as well as in *Ln. mesenteroides* M2 isolates. The *bglA* gene, was found to be present in all *Ln. cremoris* isolates, as well as *Ln. mesenteroides* M1, M3, and M4 isolates, however the gene was truncated and was identified as inactive by the Seed Viewer. A genetic potential for metabolism of trehalose was found, annotated as *treA* in *Ln. mesenteroides* and the *Ln. lactis* of dairy origin, and as *TrePP* in *Ln. pseudomesenteroides* and *Ln. lactis* 91922. Genes encoding for trehalose transport were not found in *Ln. mesenteroides* M3 and M4, indicating that these lineages are not able to metabolize trehalose from the environment. Xylose isomerase (*xylA*) and xylose kinase (*xylB*) genes were found in all *Leuconostoc* isolates, but the genes were heavily truncated in *Ln. cremoris* isolates and *Ln. mesenteroides* M3 and M4 isolates. Isolates with full length *xylA* and *xylB* genes also contained the gene *xylG*, encoding for a xylose transport protein.

TABLE 3 | Genetic potential for metabolism of carbohydrates indicated by the presence or absence of enzymes crucial to metabolism of substrates.

Gene(s)	<i>Ln. cremoris</i>			<i>Ln. mesenteroides</i>				<i>Ln. pseudomesenteroides</i>				<i>Ln. lactis</i>	
	C1 (n = 13)	C2 (n = 5)	C3 (n = 2)	M1 (n = 3)	M2 (n = 4)	M3 (n = 3)	M4 (n = 2)	P1 (n = 6)	P2 (n = 4)	P3 (n = 5)	P4 (n = 8)	L1 (n = 3)	
<i>araBAD</i>	—	—	—	+	+	—	—	—	—	—	—	—	+(33%)
<i>malP</i>	—	—	—	#	+	+	—	+	+	+	+	+	+
<i>malEFG</i>	—	—	—	—	—	—	—	#	#	#	#	—	+
<i>malX</i>	—	—	—	—	—	—	—	—	—	—	—	—	+
<i>malL</i>	—	—	—	+	+	+	—	+	+	+	+	+	+
<i>malR</i>	—	—	—	+	+	+	—	+	+	+	+	+	+
<i>lacL</i>	+	+	+	+(66%)	+(50%)	+	+	+	+	+	+	+	—
<i>lacM</i>	+	+	+	+(66%)	+(50%)	+	+	+	+	+	+	+	—
<i>lacZ</i>	#	#	#	+	+	#	#	—	—	—	—	—	+
<i>lacS</i>	#	#	#	+	+	+	+	+	+	+	+	+	+
<i>galEKT</i>	+	+	+	+	+(75%)	+	+	+	+	+	+	+	+
<i>manXYZ</i>	+	+	+	+	+	+	+	+	+	+	+	+	+
<i>manA</i>	+	+	+	+	+	+	+	+	+	+	+	+	+
<i>scrB</i>	—	—	—	+	+	+	+	—	—	—	+	+	+
<i>xylABG</i>	#	#	#	+	+	#	#	+	+	+	+	+	+
<i>treA</i>	—	—	—	+	+	+	+	—	—	—	—	—	#(66%)
<i>trePP</i>	—	—	—	—	—	—	—	+	+	+	+	+	+(33%)
<i>bglA</i>	#	#	#	#	+	#	#	+	+	+	+	+	+
<i>fruA</i>	—	—	—	—	—	—	—	—	—	—	—	—	+
<i>levE</i>	—	—	—	—	+	+	+	+	+	+	+	+	—
<i>frk</i>	#	#	#	+	+	+	+	+	+	+	+	+	+
<i>citCDEFGOS</i>	+	+	+	+	+(50%)	—	+	+	+	+	+	+	+(66%)
<i>fba</i>	—	—	—	—	—	—	—	+	+	+	+	+	+

+, gene presence. —, gene absence; #, gene present but truncated. Number in parenthesis signifies percentage of isolates where gene was present. All the isolates were able to metabolize glucose and lactose. The number given in parenthesis is given for the percentage of isolates within the lineage with the gene. Genes are abbreviated as follows: *araBAD*, arabinose metabolism pathway; *malP*, maltose phosphorylase; *malEFG*, maltose transport genes; *malX*, maltose/maltodextrin binding precursor; *malL*, sucrose-isomaltose; *malR*, maltose operon regulatory gene; *lacL*, beta-galactosidase, big subunit; *lacM*, beta-galactosidase, small subunit; *lacZ*, beta-galactosidase; *lacS*, lactose permease; *galEKT*, galactose metabolism; *manXYZ*, mannose transport genes; *manA*, mannose-6-phosphate isomerase; *scrB*, sucrose-6-phosphate hydrolase; *xylABG*, xylose isomerase, xylose kinase, xylose transport protein; *treA*, trehalose-6-phosphate hydrolase; *trePP*, trehalose-6-phosphate phosphorylase; *bglA*, beta-D-glucosidase; *fruA* and *levE*, fructose PTS; *frk*, fructokinase; *citCDEFGOS*, citrate metabolism operon; *fba*, fructose bisphosphate aldolase

(III) Citrate Metabolism

All the dairy strains in this study contained the genes necessary for uptake and metabolism of citrate. These genes are found in an operon comprised of *citC* (citrate lyase ligase), *citDEF* (citrate lyase), *citG* (holo-ACP synthase), *citO* (transcriptional regulator) and *citS* (Na⁺ dependent citrate transporter). A citrate/malate transporter annotated *cimH* was present in *Ln. mesenteroides* subspecies isolates, but was not present in any of the *Ln. lactis* or *Ln. pseudomesenteroides* isolates. In the *Ln. cremoris* and *Ln. pseudomesenteroides* genomes, the *cit* operon is flanked by two IS116/IS110/IS902 family transposases, suggesting it may have been acquired by horizontal gene transfer. In these bacteria, the operon appears to be located on the chromosome, a finding supported by the genome assembly, which organizes the *cit* operon on a contig containing a number of essential genes, and by read coverage analysis that shows a continuous gapless coverage through the contig, with no elevation in read coverage across the *cit* operon. The *citCDEFGOS* operons of *Ln. mesenteroides* and *Ln. lactis*, however, appear to be located on a plasmid, since in all cases they assembled on a contig, which includes a site of replication and not essential genes. The *cit* operon is

highly conserved in the *Ln. cremoris* and *Ln. pseudomesenteroides* genomes with >97% DNA sequence identity between all the isolates. The likely to be plasmid-encoded *cit* operon found in *Ln. mesenteroides* and *Ln. lactis* genomes is also highly conserved between the isolates (>99% identity), however it is significantly different from the chromosomally encoded *cit* operon present in *Ln. cremoris* and *Ln. pseudomesenteroides* (50–65% DNA sequence identity for each gene). None of the strains of non-dairy origin included in this study contained the citrate genes, indicating that the ability to metabolize citrate plays an important role in the successful adaption to the milk environment.

(IV) Proteolytic Activity

Leuconostocs grow in association with the lactococci in dairy fermentations, and commonly grow poorly in milk without the presence of lactococci. The general explanation for this poor growth is their lack of proteinase activity, making them dependent on small peptides from lactococcal proteinase activity. Screening all the isolates for genes involved in peptide and proteolytic activity revealed a number of differences between the lineages (Table 4). The genes encoding for the OppABCDF

TABLE 4 | Genetic potential for proteolytic activity.

Gene(s)	<i>Ln. cremoris</i>			<i>Ln. mesenteroides</i>				<i>Ln. pseudomesenteroides</i>				<i>Ln. lactis</i>	
	C1 (n = 13)	C2 (n = 5)	C3 (n = 2)	M1 (n = 3)	M2 (n = 4)	M3 (n = 3)	M4 (n = 2)	P1 (n = 6)	P2 (n = 4)	P3 (n = 5)	P4 (n = 8)	L1 (n = 3)	
<i>prtP</i>	—	—	—	+ (33%)	—	—	+	+	+	+	+	+	+ (66%)
<i>pepA</i>	+	+	+	+	+	+	+	+	+	+	+	+	+
<i>pepC</i>	+	+	+	+	+	+	+	+	+	+	+	+	—
<i>pepF</i>	+	+	+	+	+	+	+	+	+	+	+	+	+
<i>pepN</i>	—	—	—	+	+	+	+	+	+	+	+	+	+
<i>pepO</i>	+	+	+	+	+	+	+	+	+	+	+	+	+
<i>pepQ</i>	+	+	+	+	+	+	+	+	+	+	+	+	+
<i>pepS</i>	+	+	+	+	+	+	+	+	+	+	+	+	+
<i>pepT</i>	+	+	+	+	+	+	+	+	+	+	+	+	+
<i>pepV</i>	—	—	+	—	—	—	—	+	+	+	+	+	—
<i>pepX</i>	#	#	#	+	+	+	+	+	+	+	+	+	—
<i>oppABCD</i>	#	#	#	+	+	+	+	+	+	+	+	+	+

+, gene presence; —, gene absence; #, gene(s) present but truncated. Number in parenthesis indicates percentage of isolates where gene was present. Genes are abbreviated as follows: *prtP*, type-II serine proteinase; *pepA*, glutamyl aminopeptidase; *pepC*, aminopeptidase C; *pepF*, oligopeptidase; *pepN*, aminopeptidase N; *pepO*, neutral endopeptidase; *pepS*, aminopeptidase; *pepT*, peptidase T; *pepV*, beta-ala-xaa dipeptidase; *pepX*, xaa-pro dipeptidyl-peptidase; *oppABCD*, peptide ABC transporter operon.

system were found in all *Leuconostoc* genomes. However, in *Ln. cremoris* genomes, the *oppA* gene was missing, and the *oppB* gene was severely truncated. A gene encoding for a PII-type serine proteinase (PrtP) best known for its action on caseins was found in all *Ln. pseudomesenteroides* genomes, dairy *Ln. lactis* genomes, *Ln. mesenteroides* M4 and 33% of *Ln. mesenteroides* M1 genomes. All the sequenced *Leuconostoc* strains coded for a range of peptidases and aminotransferases. The *Ln. cremoris* isolates did not contain the *pepN* gene, but had the other general aminopeptidase gene, *pepC*, which was found to be missing from *Ln. lactis* genomes. The *pepX* gene, encoding for the enzyme x-prolyl dipeptidyl aminopeptidase was truncated in *Ln. cremoris* (534 amino acids) compared to the *pepX* of other *Leuconostoc* strains (778–779 amino acids). The *pepA*, *pepF*, *pepO*, *pepQ*, *pepS*, and *pepT* genes were present in all *Leuconostoc* isolates. Finally, all *Ln. pseudomesenteroides* have the *pepV* gene, encoding β-ala-dipeptidase. This dipeptidase has been shown to cleave dipeptides with an N-terminal β-Ala or D-alanine residue, such as carnosine and to a lesser extent, was shown to catalyze removal of N-terminal amino acids from a few distinct tripeptides in *Lactobacillus delbrueckii* subsp. *lactis* (Vongerichten et al., 1994).

CRISPR-Cas in *Ln. lactis* and *Ln. pseudomesenteroides*

Ln. lactis 91922 and all the *Ln. pseudomesenteroides* isolates included in this study contained CRISPR-Cas genes with repeat regions.

Composition of Leuconostocs in Starter Cultures

The *Leuconostoc* core gene library was used to devise a scheme for species and subspecies quantification in starter cultures by amplicon sequencing. Core genes were screened for sequence variation and for targeted-amplicon suitability. After curation, the top three candidates were 16S rRNA, *rpoB*, and *eno*. While the

full-length 16S rRNA sequence enables differentiation of species and subspecies, any region shorter than 500 bp is only able to differentiate between species, and then only when using the nucleotides between position 150–550, encompassing the V2 and V3 regions of 16S rRNA. However, the sequences of 16S rRNA and the *rpoB* loci were too similar to the same genes in lactococci to allow for primer design specific for leuconostocs, and thus were unsuitable for quantification of leuconostocs. The gene encoding enolase (*eno*) did allow for *Leuconostoc* specific primer design, and was used in targeted-amplicon sequencing to analyze the diversity of leuconostocs in the five starter cultures. The analysis revealed great differences between the starter cultures (Figure 4). *Ln. cremoris* dominated the *Leuconostoc* populations in cultures A, D and E, *Ln. pseudomesenteroides* was most abundant in cultures B and C. Most of the *Ln. cremoris* in cultures A and D were of the MPCA type (*Ln. cremoris* C1) unable to grow on MRS, while MRS type *Ln. cremoris* dominated in culture E (data not shown). Relatively low levels of *Ln. mesenteroides* and *Ln. dextranicum* were found in all cultures, the highest being 14% in culture B. *Ln. lactis* was only found in one of the starter cultures, culture E, where it constituted 17% of the leuconostocs.

DISCUSSION

Decades have passed since Dr. Ellen Garvie laid the foundation for the taxonomy of dairy relevant leuconostocs, and Dr. John Farrow expanded this list to include *Ln. pseudomesenteroides*. Their work has been the basis for classification of leuconostocs since then.

The *Ln. pseudomesenteroides* species was described for the first time in 1898 (Farrow et al., 1989), however its presence in a dairy starter culture was not described before 2014 (Pedersen et al., 2014b). Identification of leuconostocs by phenotypical traits or by partial 16S rRNA sequencing does not reliably distinguish between all species and misidentification has been

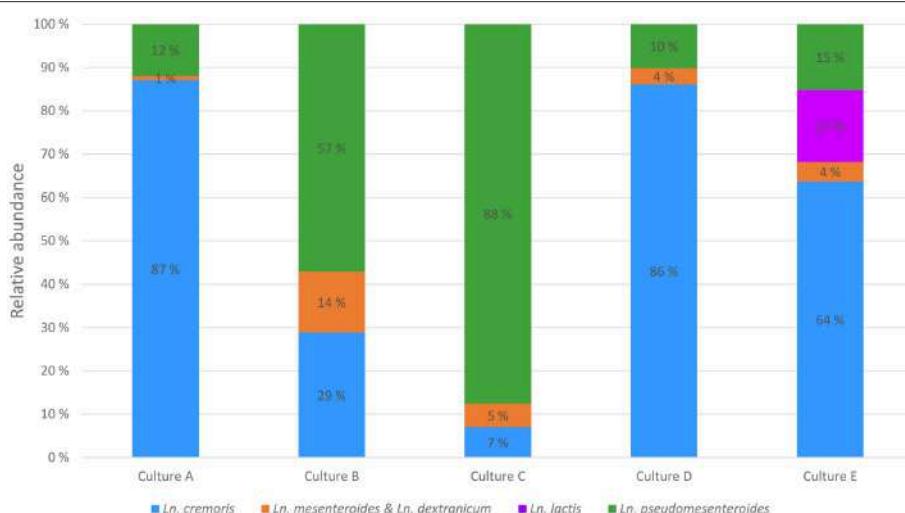


FIGURE 4 | Composition of leuconostocs in five starter cultures A–E using targeted-amplicon sequencing of the *eno* gene.

common. After genomic analysis, several isolates previously identified as *Ln. mesenteroides* subspecies proved to be *Ln. pseudomesenteroides* and isolates may have been misidentified in other studies as well. Surprisingly, the strain LbT16 (Accession No: LAYV00000000) reported to be *Ln. cremoris* by Campedelli et al. (2015) was identified as *Ln. mesenteroides* when characterized by its genomic content and its full length 16S rRNA sequence. Misidentification of *Ln. cremoris* is also uncommon. Compared to other dairy leuconostocs, *Ln. cremoris* grow slower, to a lower density and not at temperatures of 30°C or higher. In addition, a large proportion of *Ln. cremoris* type strains are not able to grow on MRS. These characteristics provide the means for reliable phenotypical identification of *Ln. cremoris*. However, phenotypical differentiation between other *Ln. mesenteroides* subspecies, *Ln. lactis* and *Ln. pseudomesenteroides* remains unreliable. In this study, dairy relevant leuconostocs are characterized using a genomics approach and the diversity of leuconostocs in five commercial DL-type starter cultures is analyzed.

The genomic analysis clearly separated leuconostocs by species, subspecies, and enabled intra-species differentiation. Interestingly, the genomic analysis did not distinguish *Ln. dextranicum* from *Ln. mesenteroides*. The strain-to-strain variation was higher than the differences between subspecies. The *dextranicum* subspecies has been previously defined by phenotypical traits only and separate subspecies distinction is not justified by the genomic data of this study. On the other hand, the pan-genomic analysis separated *Ln. mesenteroides* isolates by habitat. The dairy strains clearly differ from those isolated from plant material, the former have smaller genomes and utilize a more restricted range of carbohydrates. The two subspecies *Ln. mesenteroides* and *Ln. cremoris* share a large amount of genetic content with high identity scores, reflecting a close phylogenetic relationship. However, many genes present in *Ln. mesenteroides* are found to be truncated, contain deletions or are completely missing in *Ln. cremoris*. Adaptation of dairy

strains to the milk environment involved acquisition of the plasmid-encoded *lacLM* by horizontal gene transfer (Obst et al., 1995), which in turn permitted loss of a functional *lacZ*. Some of the dairy *Ln. mesenteroides*, and all of the *Ln. cremoris* isolates carry a deletion in the *lacZ* gene. The dairy *Ln. mesenteroides* and in particular *Ln. cremoris* display telltale signs of a prolonged degenerative evolution, likely the result of a long period of growth in milk. In this environment, the leuconostocs have evolved alongside lactococci. All the dairy strains included in this study contain the *cit* operon comprised of *citC* (citrate lyase ligase), *citDEF* (citrate lyase), *citG* (holo-ACP synthase), *citO* (transcriptional regulator) and *citS* (Na⁺ dependent citrate transporter). The *citCDEFGOS* operon organization is different from the operon in *Lactococcus lactis*, which lacks *citO* and the *citS* transporter (Drider et al., 2004). In citrate positive *Lactococcus lactis*, homologs of *citO* (*citR*) and the *citS* (*citP*) are located on a plasmid (Magni et al., 1994). The presence of the *citCDEFGOS* genes enable so-called citrolactic fermentation, co-metabolism of sugar and citrate providing the cells with higher energy yield and proton motive force (Marty-Teysset et al., 1996). In *Ln. lactis* and *Ln. mesenteroides*, this operon has been linked to a ~22-kb plasmid, inferred by phenotypical studies in combination with monitoring the presence of mobile genetic elements (Lin et al., 1991; Vaughan et al., 1995). In the study by Vaughan et al. (1995), *Ln. mesenteroides* was shown to retain its ability to metabolize citrate after losing three of its four plasmids. Moreover, after curing, a derivative isolate without the ability to degrade citrate still contained the fourth plasmid. Our data indicates that for *Ln. cremoris* and *Ln. pseudomesenteroides*, this is not the case. In all the *Ln. cremoris* and *Ln. pseudomesenteroides* genomes included in this study, the *cit* operon is located on the chromosome in a region with mobile element characteristics. A low level of genetic drift is indicated by the high sequence similarity between the *cit* operons of *Ln. cremoris* and *Ln. pseudomesenteroides* suggesting that the acquisition of these genes is quite recent,

possibly from a common donor. The chromosomally encoded *cit* operon of *Ln. cremoris* and *Ln. pseudomesenteroides* was significantly different from the highly conserved and likely to be plasmid-encoded *cit* operon found in *Ln. lactis* and *Ln. mesenteroides*. These results indicate that the plasmid encoded *cit* operon originates from a different source and time. None of the strains of non-dairy origin included in this study contained the citrate metabolism genes, indicating that the ability to metabolize citrate also plays an important role in the successful adaption to the milk environment. The manufacture of Dutch-type cheeses has been going on for centuries and the starter cultures have been maintained by so-called “back slopping” for the last one and a half century, where new milk is inoculated with whey from the previous batch. This technique for propagating starter cultures is still being used and recent studies have shown that the complex starter cultures maintain a highly stable composition with regards to lactococci (Erkus et al., 2013). Culture composition may change over a short period of time depending on growth conditions and bacteriophage predation, but the microbial community is sustained in the long run. In this study, we show a large variation in the amount and composition of the *Leuconostoc* populations in cheeses starter cultures. Three of the starter cultures (A, D, and E) were dominated by *Ln. cremoris*, and for culture A and D, the majority of these were unable to grow on MRS. The other two starter cultures (B and C) were dominated by *Ln. pseudomesenteroides*. Interestingly, the cultures dominated by *Ln. cremoris* also contain *Ln. pseudomesenteroides* strains. *Ln. pseudomesenteroides* growth rates in pure culture are significantly higher than that of *Ln. cremoris* at temperatures above 20°C, so the microbial community is preserved, either by the starter culture developers, or by the microbial community itself. Little knowledge exists on how the diversity of leuconostocs is affected by manufacturing procedures. According to Thunell (1995) and Vedamuthu (1994) the only leuconostocs relevant in dairy are *Ln. cremoris* and *Ln. lactis*, but in this study, *Ln. lactis* was detected only in culture E, which was dominated by *Ln. cremoris*. In two of the starter cultures studies in this work, *Ln. pseudomesenteroides* was the dominating *Leuconostoc*, which shows that they are highly relevant in the production of cheese. This is also reflected by recent studies, where the presence of *Ln. pseudomesenteroides* is more frequently reported (Callon et al., 2004; Porcellato and Skeie, 2016; Østlie et al., 2016). It is tempting to speculate that starter culture manufacturers have altered the conditions for culture propagation or manipulated the strain collections, thereby altering the culture dynamics between strains in favor of *Ln. pseudomesenteroides*.

The differences between the starter cultures could have an impact on the characteristics of the cheese product. *Ln. cremoris* lacks a wide range of genes involved in carbohydrate metabolism and proteolytic activity, and studies have shown that *Ln. cremoris* and *Ln. pseudomesenteroides* differ significantly in their ability to produce a wide range of volatile compounds (Pedersen et al., 2016). Most notably, the amount of acetoin and diacetyl in model-cheeses produced with only *Ln. pseudomesenteroides* was negligible. This was supported by our data, which showed that the *Ln. pseudomesenteroides* P4 isolates lack the genes necessary for

reduction of diacetyl to acetoin and 2,3-butandiol. In addition, these isolates lacked the genes *ilvB* and *ilvH* encoding acetolactate synthetase large and small subunits, which is found in all *Ln. mesenteroides* subspecies isolates. However, a different gene *alsS*, encoding the same function, was found in all leuconostocs, including *Ln. pseudomesenteroides*. Studies on α -acetolactate synthase (ALS) and α -acetolactate decarboxylase (ALDC) activity in *Ln. mesenteroides* subspecies and *Ln. lactis* showed that the activity of both ALS and ALDC was higher for *Ln. lactis* (which does not have the *ilv* or *leu* operon) than that of *Ln. cremoris* (which does have part of these two operons) (Monnet et al., 1994). For comparison, the ALS activity of *Lc. lactis* biovar *diacetylactis* was comparable or in some cases even higher than that of *Ln. lactis*. *Ln. pseudomesenteroides* was not included in the study, but data from semi-hard cheeses comparing the acetoin and diacetyl concentrations revealed lower concentrations in mock starters containing *Ln. pseudomesenteroides* compared to mock starters containing *Ln. cremoris* (Pedersen et al., 2016). This observation could be attributed to the rapid growth rate of *Ln. pseudomesenteroides* when compared to that of *Ln. cremoris*. The presence of the degenerated *ilv* and *leu* operons could somehow be negative to *Ln. cremoris* growth rate. Indeed, when cloning of the *ilv* operon into *Escherichia coli*, the presence of *Leuconostoc ilvB* was strongly detrimental to growth, while recombinant strains with an insertion in the *Leuconostoc ilvB* genes displayed normal growth. Their hypothesis was that expression of *ilvB* without a functional branched chain amino acid biosynthesis mechanism could interfere with energy metabolism via pyruvate (Cavin et al., 1999).

In dairy fermentations, the leuconostocs grow in association with the lactococci. Whether the associative growth is of mutual benefit to the leuconostocs and lactococci has not been determined. Literature often attributes the poor growth of leuconostocs to the lack of protease activity (Vedamuthu, 1994; Thunell, 1995). However, the ability to acidify milk in pure culture has been described for *Ln. pseudomesenteroides* (Cardamone et al., 2011), and we identified genetic potential for caseinolytic activity in *Ln. pseudomesenteroides* in our data. This would enable *Ln. pseudomesenteroides* to grow better in milk than *Ln. cremoris*, which lacks the capacity for protease, as well as a functional peptide uptake system due to the lack of OppA, which is responsible for the uptake of extracellular peptides. An argument for mutually beneficial growth has been made by superimposing metabolic pathways from lactococci and leuconostocs, indicating a potential for metabolic complementation between the two genera (Erkus et al., 2013). One can be forgiven for thinking *Ln. pseudomesenteroides* the better bacteria of the two based on these tidbits of information alone. However, both *Ln. cremoris* and *Ln. pseudomesenteroides* have shown to be significant to the production of cheeses. It is difficult to conclude which *Leuconostoc* species produces the highly subjective matter of the better cheese product. The concentration of volatile compounds, fatty acid derivatives, acetoin, diacetyl, and amino acid derivates in products have been shown to diverge significantly, depending on which *Leuconostoc* species is added to the mixture of lactococci (Pedersen et al., 2016).

In conclusion, the dairy-associated leuconostocs are highly adapted to grow in milk. Comparative genomic analysis reveals great differences between the *Leuconostoc* species and subspecies accustomed to the dairy environment, where they grow in association with the lactococci. The composition of the *Leuconostoc* population is significantly different between commercial starter cultures, which ultimately affects the characteristics and quality of the product. A better understanding of *Leuconostoc* microbial dynamics and the functional role of different dairy leuconostocs could be of great importance and be an applicable tool in ensuring consistent manufacture of high quality product. Currently, no detailed information on the relative amount or diversity of the *Leuconostoc* population in starter cultures is available to the industry. We provide a culture independent method for robust identification and quantification of *Leuconostoc* species in mixed microbial communities, enabling quantification of leuconostocs in starter cultures, as well as monitoring the diversity of leuconostocs through the cheese production process.

AUTHOR CONTRIBUTIONS

CF isolated and sequenced bacterial strains, performed the sequencing work in Norway (of all CF and H-isolates in addition to all amplicon sequencing), analyzed the data, wrote the R-scripts, devised the methods and wrote the manuscript.

REFERENCES

- Alegria, A., Delgado, S., Florez, A. B., and Mayo, B. (2013). Identification, typing, and functional characterization of *Leuconostoc* spp. strains from traditional, starter-free cheeses. *Dairy Sci. Technol.* 93, 657–673. doi: 10.1007/s13594-013-0128-3
- Ardö, Y., and Varming, C. (2010). Bacterial influence on characteristic flavour of cheeses made with mesophilic DL-starter. *Aust. J. Dairy Technol.* 65, 153–158.
- Auty, M. A., Gardiner, G. E., McBrearty, S. J., O'Sullivan, E. O., Mulvihill, D. M., Collins, J. K., et al. (2001). Direct *in situ* viability assessment of bacteria in probiotic dairy products using viability staining in conjunction with confocal scanning laser microscopy. *Appl. Environ. Microbiol.* 67, 420–425. doi: 10.1128/AEM.67.1.420-425.2001
- Bandell, M., Lhotte, M. E., Marty-Teyset, C., Veyrat, A., Prévost, H., Dartois, V., et al. (1998). Mechanism of the citrate transporters in carbohydrate and citrate catabolism in *Lactococcus* and *Leuconostoc* species. *Appl. Environ. Microbiol.* 64, 1594–1600.
- Barrangou, R., Yoon, S. S., Breidt, F. Jr., Fleming, H. P., and Klaenhammer, T. R. (2002). Characterization of six *Leuconostoc fallax* bacteriophages isolated from an industrial sauerkraut fermentation. *Appl. Environ. Microbiol.* 68, 5452–5458. doi: 10.1128/AEM.68.11.5452-5458.2002
- Björkroth, K. J., Geisen, R., Schillinger, U., Weiss, N., De Vos, P., Holzapfel, W. H., et al. (2000). Characterization of *Leuconostoc gasicomitatum* sp. nov., associated with spoiled raw tomato-marinated broiler meat strips packaged under modified-atmosphere conditions. *Appl. Environ. Microbiol.* 66, 3764–3772. doi: 10.1128/AEM.66.9.3764-3772.2000
- Callon, C., Millet, L., and Montel, M. C. (2004). Diversity of lactic acid bacteria isolated from AOC Salers cheese. *J. Dairy Res.* 71, 231–244. doi: 10.1017/S0022029904000159
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421
- FB, HØ, TP, HK, and HN provided bacterial isolates for a larger diversity. WK and LH performed the sequencing of isolates in Denmark. Supervision of danish activities was provided by FV. Supervision of Norwegian activities was provided by HK, HØ, and HH. All co-authors were involved in reviewing and commenting on the manuscript prior to its submission. A large contribution to final editing was made by HN and JB.
- FUNDING**
- This work was funded by the Norwegian Research Council, TINE SA, and the Danish Council for Independent Research.
- ACKNOWLEDGMENTS**
- We are grateful to TINE SA for providing culture samples and Dorota Dynda for providing isolates from Twarog. This work was funded by the Norwegian Research Council, TINE SA, and the Danish Council for Independent Research.
- SUPPLEMENTARY MATERIAL**
- The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.00132/full#supplementary-material>

- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461
- Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604
- Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575
- Erkus, O., de Jager, V. C., Spus, M., van Alen-Boerrigter, I. J., van Rijswijck, I. M., Hazelwood, L., et al. (2013). Multifactorial diversity sustains microbial community stability. *ISME J.* 7, 2126–2136. doi: 10.1038/ismej.2013.108
- Farrow, J. A. E., Facklam, R. R., and Collins, M. D. (1989). Nucleic acid homologies of some vancomycin-resistant leuconostocs and description of *Leuconostoc citreum* sp. nov. and *Leuconostoc pseudomesenteroides* sp. nov. *Int. J. Syst. Evol. Microbiol.* 39, 279–283.
- Galili, T. (2015). *Dendextend: Extending R's dendrogram functionality*. R package version 0.18.3. [Computer software]. Available online at: <http://CRAN.R-project.org/package=dendextend>
- Garvie, E. I. (1960). The genus *Leuconostoc* and its nomenclature. *J. Dairy Res.* 27, 283–292. doi: 10.1017/S0022029900010359
- Garvie, E. I. (1967). The growth factor and amino acid requirements of species of the genus *Leuconostoc*, including *Leuconostoc paramesenteroides* (sp.nov.) and *Leuconostoc oenos*. *Microbiology* 48, 439–447. doi: 10.1099/00221287-48-3-439
- Garvie, E. I. (1969). Lactic dehydrogenases of strains of the genus *Leuconostoc*. *Microbiology* 58, 85–94. doi: 10.1099/00221287-58-1-85
- Garvie, E. I. (1979). Proposal of neotype strains for *Leuconostoc mesenteroides* (Tsenkovskii) van Tieghem, *Leuconostoc dextranicum* (Beijerinck) Hucker and Pederson, and *Leuconostoc cremoris* (Knudsen and Sørensen) Garvie. *Int. J. Syst. Evol. Microbiol.* 29, 149–151.
- Garvie, E. I. (1983). NOTES: *Leuconostoc mesenteroides* subsp. *cremoris* (Knudsen and Sørensen) comb. nov. and *Leuconostoc mesenteroides* subsp. *dextranicum* (Beijerinck) comb. nov. *Int. J. Syst. Evol. Microbiol.* 33, 118–119.
- Garvie, E. I., Zezula, V., and Hill, V. A. (1974). Guanine plus cytosine content of the deoxyribonucleic acid of the *leuconostocs* and some heterofermentative lactobacilli. *Int. J. Syst. Evol. Microbiol.* 24, 248–251. doi: 10.1099/00207713-24-2-248
- Godon, J. J., Delorme, C., Bardowski, J., Chopin, M. C., Ehrlich, S. D., and Renault, P. (1993). Gene inactivation in *Lactococcus lactis*: branched-chain amino acid biosynthesis. *J. Bacteriol.* 175, 4383–4390. doi: 10.1128/jb.175.14.4383-4390.1993
- Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* 35, W52–W57. doi: 10.1093/nar/gkm360
- Gu, C. T., Wang, F., Li, C. Y., Liu, F., and Huo, G. C. (2012). *Leuconostoc mesenteroides* subsp. *sutonicum* subsp. nov. *Int. J. Syst. Evol. Microbiol.* 62, 1548–1551. doi: 10.1099/ijss.0.031203-0
- Gunnewijk, M. G., and Poolman, B. (2000). HPr(His approximately P)-mediated phosphorylation differently affects counterflow and proton motive force-driven uptake via the lactose transport protein of *Streptococcus thermophilus*. *J. Biol. Chem.* 275, 34080–34085. doi: 10.1074/jbc.M003513200
- Hache, C., Cachon, R., Wache, Y., Belguendouz, T., Riondet, C., Deraadt, A., et al. (1999). Influence of lactose-citrate co-metabolism on the differences of growth and energetics in *Leuconostoc lactis*, *Leuconostoc mesenteroides* ssp. *mesenteroides* and *Leuconostoc mesenteroides* ssp. *cremoris*. *Syst. Appl. Microbiol.* 22, 507–513. doi: 10.1016/S0723-2020(99)80002-2
- Hemme, D., and Foucaud-Scheunemann, C. (2004). *Leuconostoc*, characteristics, use in dairy technology and prospects in functional foods. *Int. Dairy J.* 14, 467–494. doi: 10.1016/j.idairyj.2003.10.005
- Hugenholtz, J. (1993). Citrate metabolism in lactic acid bacteria. *FEMS Microbiol. Rev.* 12, 165–178. doi: 10.1111/j.1574-6976.1993.tb00017.x
- Johansen, E., and Kibbenich, A. (1992). Characterization of *Leuconostoc* isolates from commercial mixed strain mesophilic starter cultures. *J. Dairy Sci.* 75, 1186–1191. doi: 10.3168/jds.S0022-0302(92)77865-5
- Jung, J. Y., Lee, S. H., Lee, S. H., and Jeon, C. O. (2012). Complete genome sequence of *Leuconostoc mesenteroides* subsp. *mesenteroides* strain J18, isolated from kimchi. *J. Bacteriol.* 194, 730–731. doi: 10.1128/JB.06498-11
- Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199
- Kleppen, H. P., Nes, I. F., and Holo, H. (2012). Characterization of a *Leuconostoc* bacteriophage infecting flavor producers of cheese starter cultures. *Appl. Environ. Microbiol.* 78, 6769–6772. doi: 10.1128/AEM.00562-12
- Lin, J., Schmitt, P., and Diviès, C. (1991). Characterization of a citrate-negative mutant of *Leuconostoc mesenteroides* subsp. *mesenteroides*: metabolic and plasmidic properties. *Appl. Microbiol. Biotechnol.* 34, 628–631.
- Locdics, T. A., and Steenson, L. R. (1990). Characterization of bacteriophages and bacteria indigenous to a mixed-strain cheese starter. *J. Dairy Sci.* 73, 2685–2696. doi: 10.3168/jds.S0022-0302(90)78953-9
- Magni, C., de Felipe, F. L., Sesma, F., López, P., and de Mendoza, D. (1994). Citrate transport in *Lactococcus lactis* biovar *diacetylactis*: expression of the plasmid-borne citrate permease p. *FEMS Microbiol. Lett.* 118, 75–82. doi: 10.1111/j.1574-6968.1994.tb06806.x
- Marty-Teyset, C., Posthuma, C., Lolkema, J. S., Schmitt, P., Divies, C., and Konings, W. N. (1996). Proton motive force generation by citrolactic fermentation in *Leuconostoc mesenteroides*. *J. Bacteriol.* 178, 2178–2185. doi: 10.1128/jb.178.8.2178-2185.1996
- Meslier, V., Loux, V., and Renault, P. (2012). Genome sequence of *Leuconostoc pseudomesenteroides* strain 4882, isolated from a dairy starter culture. *J. Bacteriol.* 194, 6637. doi: 10.1128/jb.01696-12
- Monnet, C., Phalip, V., Schmitt, P., and Diviès, C. (1994). Comparison of α -acetolactate synthase and α -acetolactate decarboxylase in *Lactococcus* spp. and *Leuconostoc* spp. *Biotechnol. Lett.* 16, 257–262. doi: 10.1007/BF00134622
- Nieto-Arribas, P., Seséña, S., Poveda, J. M., Palop, L., and Cabezas, L. (2010). Genotypic and technological characterization of *Leuconostoc* isolates to be used as adjunct starters in Manchego cheese manufacture. *Food Microbiol.* 27, 85–93. doi: 10.1016/j.fm.2009.08.006
- Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., et al. (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* 20, 714–737. doi: 10.1089/cmb.2013.0084
- Obst, M., Meding, E. R., Vogel, R. F., and Hammes, W. P. (1995). Two genes encoding the β -galactosidase of *Lactobacillus sakei*. *Microbiology* 141(Pt 12), 3059–3066. doi: 10.1099/13500872-141-12-3059
- Olsen, K. N., Brockmann, E., and Molin, S. (2007). Quantification of *Leuconostoc* populations in mixed dairy starter cultures using fluorescence in situ hybridization. *J. Appl. Microbiol.* 103, 855–863. doi: 10.1111/j.1365-2672.2007.03298.x
- Østlie, H. M., Kræggerud, H., Longva, A. B., and Abrahamsen, R. K. (2016). Characterisation of the microflora during ripening of a Norwegian semi-hard cheese with adjunct culture of propionic acid bacteria. *Int. Dairy J.* 54, 43–49. doi: 10.1016/j.idairyj.2015.10.005
- Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* 42, D206–D214. doi: 10.1093/nar/gkt1226
- Pedersen, T. B., Kot, W. P., Hansen, L. H., Sørensen, S. J., Broadbent, J. R., Vogensen, F. K., et al. (2014a). Genome sequence of *Leuconostoc mesenteroides* subsp. *cremoris* Strain T26, isolated from mesophilic undefined cheese starter. *Genome Announc.* 2:e00485-14. doi: 10.1128/genomeA.00485-14
- Pedersen, T. B., Kot, W. P., Hansen, L. H., Sørensen, S. J., Broadbent, J. R., Vogensen, F. K., et al. (2014b). Genome sequences of two *Leuconostoc pseudomesenteroides* strains isolated from danish dairy starter cultures. *Genome Announc.* 2:e00484-14. doi: 10.1128/genomeA.00484-14
- Pedersen, T. B., Vogensen, F. K., and Ardo, Y. (2016). Effect of heterofermentative lactic acid bacteria of DL-starters in initial ripening of semi-hard cheese. *Int. Dairy J.* 57, 72–79. doi: 10.1016/j.idairyj.2016.02.041
- Pérez, G., Cardell, E., and Zárate, V. (2002). Random amplified polymorphic DNA analysis for differentiation of *Leuconostoc mesenteroides* subspecies isolated from Tenerife cheese. *Lett. Appl. Microbiol.* 34, 82–85. doi: 10.1046/j.1472-765x.2002.01050.x
- Porcellato, D., Østlie, H. M., Brede, M. E., Martinovic, A., and Skeie, S. B. (2013). Dynamics of starter, adjunct non-starter lactic acid bacteria and propionic acid

- bacteria in low-fat and full-fat Dutch-type cheese. *Int. Dairy J.* 33, 104–111. doi: 10.1016/j.idairyj.2013.01.007
- Porcellato, D., and Skeie, S. B. (2016). Bacterial dynamics and functional analysis of microbial metagenomes during ripening of Dutch-type cheese. *Int. Dairy J.* 61, 182–188. doi: 10.1016/j.idairyj.2016.05.005
- Salvador, S., and Chan, P. (2004). “Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms,” in *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence* (Boca Raton, FL: IEEE Computer Society).
- Sánchez, J. I., Martínez, B., and Rodríguez, A. (2005). Rational selection of *Leuconostoc* strains for mixed starters based on the physiological biodiversity found in raw milk fermentations. *Int. J. Food Microbiol.* 105, 377–387. doi: 10.1016/j.ijfoodmicro.2005.04.025
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153
- Thunell, R. K. (1995). Taxonomy of the *Leuconostocs*. *J. Dairy Sci.* 78, 2514–2522. doi: 10.3168/jds.S0022-0302(95)76881-3
- Vaughan, E. E., David, S., Harrington, A., Daly, C., Fitzgerald, G. F., and De Vos, W. M. (1995). Characterization of plasmid-encoded citrate permease (*citP*) genes from *Leuconostoc* species reveals high sequence conservation with the *Lactococcus lactis* *citP* gene. *Appl. Environ. Microbiol.* 61, 3172–3176.
- Vedamuthu, E. R. (1994). The dairy *Leuconostoc*: use in dairy products. *J. Dairy Sci.* 77, 2725–2737. doi: 10.3168/jds.S0022-0302(94)77215-5
- Vihavainen, E. J., and Björkroth, K. J. (2009). Diversity of *Leuconostoc gasicomitatum* associated with meat spoilage. *Int. J. Food Microbiol.* 136, 32–36. doi: 10.1016/j.ijfoodmicro.2009.09.010
- Villani, F., Moschetti, G., Blaiotta, G., and Coppola, S. (1997). Characterization of strains of *Leuconostoc mesenteroides* by analysis of soluble whole-cell protein pattern, DNA fingerprinting and restriction of ribosomal DNA. *J. Appl. Microbiol.* 82, 578–588. doi: 10.1111/j.1365-2672.1997.tb03588.x
- Vogensen, F. K., Karst, T., Larsen, J. J., Krügelum, B., Ellekjaer, D., and Waagner Nielsen, E. (1987). Improved direct differentiation between *Leuconostoc cremoris*, *Streptococcus lactis* ssp. *diacetylactis*, and *Streptococcus cremoris*/*Streptococcus lactis* on agar. *Milchwissenschaft* 42, 646–648.
- Vongerichten, K. F., Klein, J. R., Matern, H., and Plapp, R. (1994). Cloning and nucleotide sequence analysis of *pepV*, a carnosinase gene from *Lactobacillus delbrueckii* subsp. *lactis* DSM 7290, and partial characterization of the enzyme. *Microbiology* 140, 2591–2600. doi: 10.1099/00221287-140-1-2591
- Ward, D. M., Weller, R., and Bateson, M. M. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* 345, 63–65.
- Warnes, G. R., Bolker, B., Bonebakker, L., R. G., Liaw, W. H. A., Lumley, T., Maechler, M., et al. (2015). *gplots: Various R Programming Tools for Plotting Data*. R package version 2.16.0. [Computer software]. Available online at: <http://CRAN.R-project.org/package=gplots>
- Wright, E. (2015). *DECIPHER: Database Enabled Code for Ideal Probe Hybridization Employing*. R package v1.16.1.
- Zeller-Péronnet, V., Brockmann, E., Pavlovic, M., Timke, M., Busch, U., and Huber, I. (2013). Potential and limitations of MALDI-TOF MS for discrimination within the species *Leuconostoc mesenteroides* and *Leuconostoc pseudomesenteroides*. *J. für Verbraucherschutz und Lebensmittelsicherheit* 8, 205–214. doi: 10.1007/s00003-013-0826-z
- Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30, 614–620. doi: 10.1093/bioinformatics/btt593
- Zhang, W., Liu, W., Song, Y., Xu, H., Menghe, B., Zhang, H., et al. (2015). Multilocus sequence typing of a dairy-associated *Leuconostoc mesenteroides* population reveals clonal structure with intragenic homologous recombination. *J. Dairy Sci.* 98, 2284–2293. doi: 10.3168/jds.2014-9227
- Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., et al. (2014). PanGP: a tool for quickly analyzing bacterial pan genome profile. *Bioinformatics* 30, 1297–1299. doi: 10.1093/bioinformatics/btu017

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Frantzen, Kot, Pedersen, Ardö, Broadbent, Neve, Hansen, Dal Bello, Østlie, Kleppen, Vogensen and Holo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Comparative Genomic Analysis Reveals Ecological Differentiation in the Genus *Carnobacterium*

Christelle F. Iskandar¹, Frédéric Borges^{1*}, Bernard Taminiau², Georges Daube², Monique Zagorec³, Benoît Remenant^{3†}, Jørgen J. Leisner⁴, Martin A. Hansen⁵, Søren J. Sørensen⁵, Cécile Mangavel¹, Catherine Cailliez-Grimal¹ and Anne-Marie Revol-Junelles¹

OPEN ACCESS

Edited by:

Jennifer Ronholm,
McGill University, Canada

Reviewed by:

Cristian Botta,
University of Turin, Italy
Rolf Dieter Joerger,
University of Delaware, USA

*Correspondence:

Frédéric Borges
frédéric.borges@univ-lorraine.fr

†Present address:

Benoît Remenant,
Laboratoire de la Santé des Végétaux,
Agence Nationale de Sécurité
Sanitaire de l'Alimentation,
de l'Environnement et du Travail,
Angers, France

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 29 November 2016

Accepted: 21 February 2017

Published: 08 March 2017

Citation:

Iskandar CF, Borges F, Taminiau B,
Daube G, Zagorec M, Remenant B,
Leisner JJ, Hansen MA,
Sørensen SJ, Mangavel C,
Cailliez-Grimal C and
Revol-Junelles A-M (2017)

Comparative Genomic Analysis
Reveals Ecological Differentiation
in the Genus *Carnobacterium*.

Front. Microbiol. 8:357.
doi: 10.3389/fmicb.2017.00357

¹ Laboratoire d'Ingénierie des Biomolécules, École Nationale Supérieure d'Agronomie et des Industries Alimentaires – Université de Lorraine, Vandoeuvre-lès-Nancy, France, ² Laboratory of Food Microbiology, Department of Food Science, Fundamental and Applied Research for Animal and Health, University of Liège, Liège, Belgium, ³ UMR1014 SECALIM, INRA, Oniris, Nantes, France, ⁴ Department of Veterinary Disease Biology, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark, ⁵ Molecular Microbial Ecology Group, University of Copenhagen, Copenhagen, Denmark

Lactic acid bacteria (LAB) differ in their ability to colonize food and animal-associated habitats: while some species are specialized and colonize a limited number of habitats, others are generalists and are able to colonize multiple animal-linked habitats. In the current study, *Carnobacterium* was used as a model genus to elucidate the genetic basis of these colonization differences. Analyses of 16S rRNA gene meta-barcode data showed that *C. maltaromaticum* followed by *C. divergens* are the most prevalent species in foods derived from animals (meat, fish, dairy products), and in the gut. According to phylogenetic analyses, these two animal-adapted species belong to one of two deeply branched lineages. The second lineage contains species isolated from habitats where contact with animal is rare. Genome analyses revealed that members of the animal-adapted lineage harbor a larger secretome than members of the other lineage. The predicted cell-surface proteome is highly diversified in *C. maltaromaticum* and *C. divergens* with genes involved in adaptation to the animal milieu such as those encoding biopolymer hydrolytic enzymes, a heme uptake system, and biopolymer-binding adhesins. These species also exhibit genes for gut adaptation and respiration. In contrast, *Carnobacterium* species belonging to the second lineage encode a poorly diversified cell-surface proteome, lack genes for gut adaptation and are unable to respire. These results shed light on the important genomic traits required for adaptation to animal-linked habitats in generalist *Carnobacterium*.

Keywords: lactic acid bacteria, *Carnobacterium*, 16S meta-barcode, comparative genomic analyses, ecological differentiation

INTRODUCTION

Lactic acid bacteria (LAB) include various genera and many species that have been investigated for decades because of their major role in food fermentations and their health benefit potential as probiotics. Understanding how LAB fitness can increase through evolution in these various habitats is critical for exploiting their beneficial properties. LAB are well-known for their ability

to colonize animal-derived food, i.e., meat, fish, and dairy products, and for being members of the gastrointestinal (GI) tract and the vagina microbiota (Douglas and Klaenhammer, 2010; Douillard and de Vos, 2014). LAB can be divided into specialized and generalist bacteria. Typically, the specialists that are used as starter cultures for a narrow range of fermented products are characterized by a low genetic diversity (Delorme et al., 2010). Their genomes exhibit signs of massive losses of genes involved in biosynthetic pathways (Douglas and Klaenhammer, 2010). To compensate the loss of these functions, genes encoding transporters for amino acids or carbohydrates were gained to allow growth in nutritional rich fermentation environments (Lorca et al., 2007). These genomic changes were accompanied by a specialization to food matrices, particularly exemplified in dairy strains. Other LAB also exhibit genomes characteristic of their ecological specialization. *Lactobacillus iners* is described to solely colonize the vaginal cavity and harbors one of the smallest LAB genomes presumably because its niche specialization allowed a substantial genome reduction (Macklaim et al., 2011; Mendes-Soares et al., 2014). Another example is the GI tract symbiont *Lactobacillus reuteri*, which is characterized by different lineages each one being apparently adapted to one particular vertebrate host: rodent or human (Frese et al., 2011). By contrast, some species are more generalist in their ability to colonize various environments and are therefore ubiquitous. Species from the genus *Enterococcus* can be found in the GI tract of animals and in a multitude of fermented foods (Santagati et al., 2012). Their adaptation to various environments is strongly linked to the presence in their genome of DNA acquired through horizontal gene transfer (HGT), resulting in a large pan genome (Santagati et al., 2012). Similarly, the genomes of some strains of the ubiquitous species *Lactobacillus rhamnosus* encode multiple lifestyle traits allowing them to reside in diverse habitats (Douillard et al., 2013). However, the interconnection between the ecology of these bacteria and their genomics is not fully understood and need further investigation. Importantly, the genomic traits responsible for adaptation to multiple animal-associated habitats are not clearly defined.

The LAB genera *Enterococcus*, *Lactobacillus*, *Lactococcus*, and *Streptococcus* have been intensively investigated, but lately, other genera including *Carnobacterium* have also drawn attention since 16S meta-barcoding studies have shown their significance in food (Chaillou et al., 2015; Duan et al., 2016; Fougy et al., 2016; Jääskeläinen et al., 2016). The genus *Carnobacterium* encompasses 11 species that have been isolated from cold and temperate environments, and from the GI tract of animals as well as from foods of animal origin such as seafood, meat, and dairy products. They are mesophilic and some species are psychrotolerant and able to grow down to 0°C. Some are halotolerant and able to grow with 8% NaCl and some are alkaliphilic with growth up to pH 9.5 (Cailliez-Grimal et al., 2014; Pikuta, 2014; Pikuta and Hoover, 2014). These traits could explain their wide distribution. However, it is apparent that some heterogeneity exists within the genus regarding habitat associations. Some species can be isolated from environments where contact with animals is likely rare

as exemplified by *Carnobacterium inhibens* subsp. *gilichinskyi* WN1359, *Carnobacterium* sp. 17-4, and *Carnobacterium* AT7 which were isolated from Siberian permafrost, sea-ice from permanently cold fjords of the Arctic Ocean, and an oceanic trench, respectively (Lauro et al., 2007; Voget et al., 2011; Leonard et al., 2013; Nicholson et al., 2015). Other species, including *C. maltaromaticum* and *C. divergens*, have been found in animal-associated habitats. These two species are the most frequently isolated carnobacteria from various sources (Leisner et al., 2007) and belong to the dominant bacterial communities in meat and fish derived food (Chaillou et al., 2015; Duan et al., 2016; Fougy et al., 2016; Jääskeläinen et al., 2016). These bacteria are therefore very interesting models to investigate the genomic traits responsible for adaptation to animal-associated habitats. In the genus *Carnobacterium*, the genome size ranges from 2.4 Mbp for *C. inhibens* subsp. *gilichinskyi* WN1359 (Leonard et al., 2013), to 3.7 Mbp for *C. maltaromaticum* (Cailliez-Grimal et al., 2013). It has been suggested that the larger genome of the latter species is the basis of its success in colonizing various habitats (Leisner et al., 2007). However, it appears that genome size is not necessarily a predictor of the ability to occupy diverse habitats since the genome of *C. divergens* strains is relatively small (~2.7 Mbp), but this species is still able to colonize various habitats (Sun et al., 2015; Remenant et al., 2016).

The aim of this study was to investigate the ecological niches *Carnobacterium* species can occupy and to identify the genomic traits responsible for the high ecological success of some *Carnobacterium* species and the possible underlying adaptive mechanisms. For that purpose, we analyzed the relative abundance of *Carnobacterium* species using 16S rDNA metagenomic data. Subsequently, we compared the genomes of *Carnobacterium* strains isolated from different environments (mainly from cold aquatic habitats), and animal-associated habitats (live animal and foods).

MATERIALS AND METHODS

16S Meta-Barcoding Sequence Analysis

The data obtained from 681 samples of various ecological origins were analyzed with a focus on *Carnobacterium*. A database of V1–V3 16S rRNA gene sequence datasets available at the FARAH Institute (University of Liège, Belgium) was used to delineate the species composition of the *Carnobacterium* genus within four types of matrices: food products, animal samples, human and animal feces and environment. This database was built from 2010 to 2015 by merging the data obtained from several single projects hosted at the FARAH Institute. The datasets were produced as previously described (Rodriguez et al., 2015) by sequencing the V1–V3 16S rDNA hypervariable region with an MiSeq sequencer using v3 reagents (ILLUMINA, USA).

Sequence read processing was employed as previously described (Rodriguez et al., 2015) using the MOTHUR software package v1.35 (Schloss and Handelsman, 2003) and the UCHIME algorithm (Edgar et al., 2011) for alignment and OTU clustering (distance 0.03) and chimera detection, respectively. 16S gene sequence reference alignment and taxonomical assignation were

based upon the SILVA database (v1.15) of full-length 16S rDNA sequences.

For each sample, Operational Taxonomic Units (OTUs) belonging to the *Carnobacterium* genus were extracted. Corresponding reads were further assigned to *Carnobacterium* species using a local BLASTn algorithm vs. the SILVA database (v1.15). Reads were assigned to a defined species when identical to the best hit (four mismatches were allowed).

Statistical differences of the different species proportion inside each type of matrix were assessed with non-parametric Kruskal-Wallis test with Dunn's *post hoc* tests using PRISM6 (GraphPad Software).

Carnobacterium Genome Analysis

The *Carnobacterium* genome sequences available at the start of this study included five *C. maltaromaticum*, one *C. inhibens* subsp. *gilichinskyi*, one *C. divergens*, and two *Carnobacterium* sp. genomes among which three were complete (*C. maltaromaticum* LMA28, *Carnobacterium* sp. 17.4, and *C. inhibens* subsp. *gilichinskyi* WN1359). The strains originated from different ecological habitats, some from animal-derived food such as the *C. maltaromaticum* strains and *C. divergens* V41, and others from environmental samples (Table 1).

The genome sequences were integrated in the MicroScope platform (Vallenet et al., 2013) to perform automatic and expert annotation of the genes, as well as comparative analysis and secretome prediction by using the integrated SignalP software (Petersen et al., 2011). The gene phyloprofile tool interface was used for searching common or specific genes/regions between a query genome and other genomes or replicons chosen from the ones available in Prokaryotic Genome DataBase (PkGDB; i.e., (re)annotation of bacterial genomes) or complete proteome downloaded from the RefSeq/WGS sections.

The pan/core genome was calculated using two methods. The pan/core genome tool accessible in the comparative genomics section was used with MICFAM parameters of 50 or 80% amino acid (aa) sequence identity and 80% coverage. The Phyloprofile tool from the MicroScope platform was used with a cut-off of 70% aa identity and 80% coverage with the best Bidirectional Best Hit (BBH).

Synteny, defined as an orthologous gene set having the same local organization in species A and B, was determined as sequence similarity by BlastP BBH with at least 30% identity on 80% of the shortest sequence (minLrap 0.8) analyses and co-localization. Metabolic pathways were predicted using the Kyoto Encyclopedia of Genes and the Genomes (KEGG) resources (Kanehisa and Goto, 2000; Kanehisa et al., 2014) and the MetaCyc database (Caspi et al., 2014). A percentage of 70% minimum identity was used to detect the specific and common genes for *C. maltaromaticum*, excluding genes with 30% minimum identity with the four other *Carnobacterium* strains.

Neighbor-joining-based phylogenetic reconstruction was based on the nucleic acid sequence of 10 housekeeping genes (*dnaK*, *gyrA*, *polA*, *lepA*, *dnaB*, *gyrB*, *secA*, *ftsZ*, *recG*, *ileS*) and was performed using MEGA6 by using the Kimura two-parameter model, including transitions and transversions. The candidate tree was tested with 1,000 bootstrap replications (Tamura et al.,

2013). The Sequence Type (MLST) was updated using e-BURST analysis from Rahman et al. (2014a). The resulting tree was rooted using the closely related species *Enterococcus faecalis* as outgroup.

The search for prophages was conducted with the PHAST Search Tool (Zhou et al., 2011).

Data Availability

The annotations were deposited at DDBJ/EMBL/GenBank under the following references: PRJEB9002 for *C. maltaromaticum* ML.1.97, and PRJEB8756 for *C. maltaromaticum* 3.18. The annotations are publicly available for consultation in MicroScope¹.

RESULTS AND DISCUSSION

Prevalence of Different *Carnobacterium* Species in Various Ecological Niches

Metagenomic data for genes encoding 16S rDNA from 681 samples were analyzed with a focus on the genus *Carnobacterium*. The samples were categorized as animal-derived food, animal organs, feces, and environment. Overall, the presence of representatives of the species *C. divergens*, *C. iners*, *C. inhibens*, *C. jeotgali*, *C. maltaromaticum*, *C. mobile*, *C. viridans*, and uncultured *Carnobacterium* sp was detected and their relative abundance is presented in Figure 1. The most abundant species was *C. maltaromaticum* accounting for 28–60% of *Carnobacterium* reads, followed by *C. divergens* (15–49%) and *Carnobacterium* spp. from lineages that have not yet been cultured and characterized (14–47% of *Carnobacterium* reads, depending on the sample origin). *C. inhibens*, *C. mobile*, and *C. viridans* accounted for lesser reads. *C. jeotgali* was detected only in environmental samples. *C. viridans* was detected only in food samples whereas *C. maltaromaticum* reads were observed in the samples from all origins. Other species were detected in the samples from two or three different habitats.

The highest species diversity was observed in food and organs (dog lungs, pig nymphal nodes, pig stomach, cattle forestomach, and cattle spleen), with a maximum of eight species in food. The lowest diversity was observed in feces where only two species were recorded. In food and organs, the most abundant species were *C. maltaromaticum*, *C. divergens*, and uncultured *Carnobacterium* sp. The OTUs assigned to the species *C. maltaromaticum* accounted for 54% and 60% of the reads associated to the genus *Carnobacterium* in food and organs, respectively (Figure 1). Interestingly, in the feces only *C. maltaromaticum* and *C. divergens* sequences were detected, each accounting for about half of the reads. In contrast, *C. maltaromaticum* sequences represented 27% of the reads, while 47% of the reads were attributed to uncultured *Carnobacterium* sp. in environmental samples. *C. divergens* was not found in samples originating from the environment.

These results strongly suggest that *C. maltaromaticum* and to a lesser extend *C. divergens* are the most prevalent species of

¹<http://www.genoscope.cns.fr/agc/microscope/home/>

TABLE 1 | Characteristics of *Carnobacterium* strains.

Species	Strain name	Origin and reference	Genome reference	Accession number
<i>C. divergens</i>	V41	Fish viscera (Pilet et al., 1994)	Remenant et al., 2016	FLLU01000001 to FLLU01000032
<i>C. inhibens</i> subsp. <i>gilichinskyi</i>	WN1359	Siberian permafrost (Leonard et al., 2013)	Leonard et al., 2013	CP006812 to CP006817
<i>C. maltaromaticum</i>	ATCC35586*	Diseased trout (Hiiu et al., 1984)	Leisner et al., 2012	NZ_AGNS00000000.1
	LMA28*	Soft ripened cheese (Millière et al., 1994)	Caillez-Grimal et al., 2013	HE999757.2
	DSM20342 MX5*	Milk with malty flavor (Miller et al., 1974)		NZ_JQMX00000000.1
	3.18*	Pork meat product (Laursen et al., 2005)	Iskandar et al., 2016	PRJEB8756
	ML.1.97*	Fresh salmon (Laursen et al., 2005)	Iskandar et al., 2016	PRJEB9002
<i>Carnobacterium</i> sp.	AT7	Aleutian trench (Lauro et al., 2007)	Lauro et al., 2007	NZ_ABHH00000000.1
	17.4	Cold seawater (Voget et al., 2011)	Voget et al., 2011	NC_015390.1, NC_015391.1

*Supplementary Figure S1.

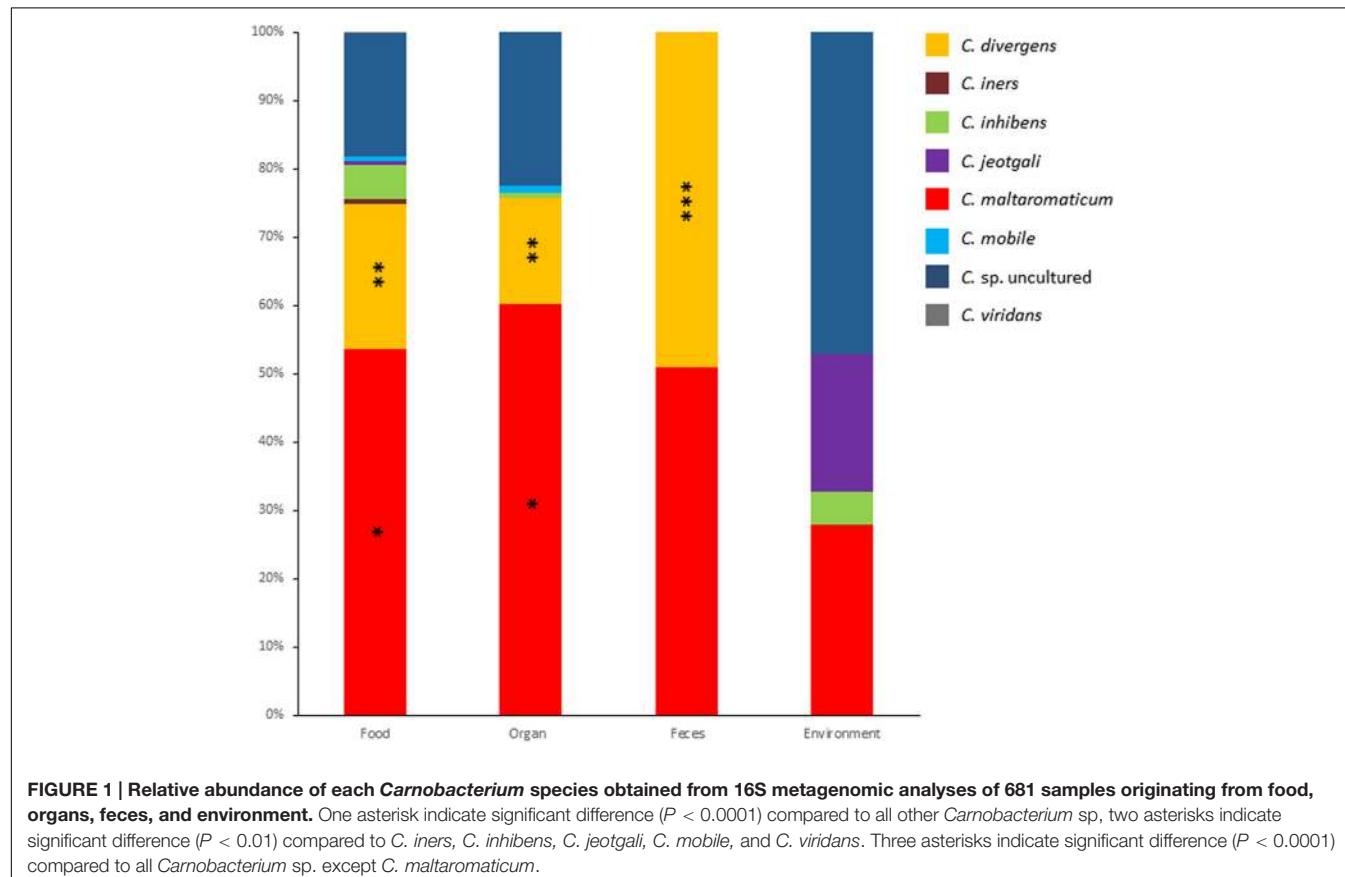


FIGURE 1 | Relative abundance of each *Carnobacterium* species obtained from 16S metagenomic analyses of 681 samples originating from food, organs, feces, and environment. One asterisk indicate significant difference ($P < 0.0001$) compared to all other *Carnobacterium* sp, two asterisks indicate significant difference ($P < 0.01$) compared to *C. iners*, *C. inhibens*, *C. jeotgali*, *C. mobile*, and *C. viridans*. Three asterisks indicate significant difference ($P < 0.0001$) compared to all *Carnobacterium* sp. except *C. maltaromaticum*.

this genus in habitats associated to animals. Although animal-associated habitats differ from each other, they also share common properties: they are nutrient-rich environments, they can induce similar stresses, they are associated with a dense and diversified microbiota and carbon is mainly available in the form of polymeric biomacromolecules. It can therefore be expected that bacteria, such as *Carnobacterium* species that are associated to animal habitats, share general properties and thereby highly contrast with other *Carnobacterium* species associated with the external environment.

Comparison of the General Genomic Features

In order to identify the adaptation factors responsible for the high ability of *C. maltaromaticum* and *C. divergens* to colonize multiple ecological niches including animal-associated habitats, a genome based analysis was conducted on nine *Carnobacterium* genomes (Table 1). The genome sizes of the five *C. maltaromaticum* strains were the largest, ranging from 3.33 to 3.87 Mbp; the other *Carnobacterium* genomes were smaller (2.35–2.74 Mbp), with the smallest being that of

C. inhibens subsp. *gilichinskyi* WN1359 (**Table 2**). Accordingly, the number of predicted CDS in each genome ranged from 3,368 to 3,812 for *C. maltaromaticum*, and from 2,268 to 2,633 for the other *Carnobacterium* (**Table 2**). *C. maltaromaticum* genomes harbored a lower GC% (34.4–34.5) than other genomes (35.2–35.3).

A phylogenetic tree based on 10 housekeeping genes was constructed. It shows that *C. maltaromaticum* and *C. divergens* on one hand, and *Carnobacterium* sp. 17.4, *Carnobacterium* sp. AT7, and *C. inhibens* subsp. *gilichinskyi* WN1359 on the other hand, are closely related. They form two monophyletic taxons sharing a common ancestor, as shown by the outgroup *E. faecalis* V583 (**Figure 2**).

The pan/core genome analysis of the nine *Carnobacterium* (cut-off of 50% aa identity and 80% coverage) showed that the core genome represents 1,130 (31%) of the predicted CDS. When comparison was restricted to *C. inhibens* subsp. *gilichinskyi* WN1359, *Carnobacterium* sp. 17-4, and *Carnobacterium* sp. AT7, the core genome increased up to 1,745–1,785 (67–73%) of the predicted CDS. Similarly, when the comparison was restricted to the 5 *C. maltaromaticum* strains and *C. divergens* V41, the core genome increased up to 1,723–1,825 (68–76%) of the predicted CDS. This result is in agreement with the phylogenetic tree showing close phylogenetic proximity of *C. maltaromaticum* and *C. divergens* on the one hand, and of *C. inhibens* subsp. *gilichinskyi* WN1359, *Carnobacterium* sp. 17-4, and *Carnobacterium* sp. AT7 on the other hand.

To be more restrictive, the pan/core genome analysis was performed with a cut-off of 70% aa identity and 80% coverage. It revealed that the strains of *C. maltaromaticum* share 2,665 CDS, and each of the five strains possesses between 279 and 644 specific genes showing strain-to-strain variations.

The gene repertoire of the different genomes was then compared to search for functions that differ between strains in order to identify the genomic traits that could explain the potential adaptation of *C. maltaromaticum* and *C. divergens* to animal-linked habitats.

CRISPR-cas and Prophages

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs; Barrangou et al., 2007), and the CRISPR associated (cas) genes confer resistance to phage infection. The role and mechanism of the CRISPR-cas system in bacterial species have been extensively studied and indicate that the spacer sequences can be considered as a signature of past exposure to exogenous DNA. Among all *Carnobacterium* genomes, only *C. divergens* V41 possessed a CRISPR-cas system. More precisely, two loci were predicted in the genome of this strain. The presence of these genes suggests that *C. divergens* V41 may have systems acting as barriers against HGT, thereby that would limit genome expansion in this taxon.

One to four prophages and/or prophage remnants were detected in all analyzed genomes except in those of *C. inhibens* subsp. *gilichinskyi* and *Carnobacterium* sp. 17.4 (**Table 2**). Their size, ranged from 11.5 to 74.4 kbp (data not shown). More importantly, complete prophages were found only in the *C. maltaromaticum* (**Table 2**).

The genomes of all *Carnobacterium* species but one – *C. maltaromaticum*- exhibit a genome of small or relatively small sizes. This suggests that the common ancestor of *Carnobacterium* exhibited a small size and that the ancestor of the species *C. maltaromaticum* experienced a massive gain of genes. Compared to *C. divergens*, no CRISPR-Cas systems were found in the genome of *C. maltaromaticum*. CRISPR-Cas provide an adaptive immunity against foreign DNA and is considered as a barrier against horizontal transfer (Barrangou et al., 2007). The lack of such a barrier against DNA transfer could have favored the acquisition of genes in the *C. maltaromaticum* lineage. Consistently, complete prophages were found in the genomes of *C. maltaromaticum* while only remnants prophages were found in the genome of *C. divergens*. Similarly, *Lactobacillus* genomes devoid of CRISPR-Cas systems exhibited the trend of being more abundant in phage sequences (Sun et al., 2015). LAB are mainly described as evolving by massive gene loss. The lineage *C. maltaromaticum* suggests that LAB may also evolve by a massive gene gain. Consistently, other ubiquitous LAB also exhibited large genomes that could be a result of similar mechanisms (Sun et al., 2015).

Secretome

Homologs of genes encoding the general secretion route (Sec-pathway) were present in all strains, while no homolog of the Twin-arginine translocation pathway (Tat-pathway) was found (data not shown). The secretome was therefore predicted by identifying genes suspected to encode signal peptide-containing proteins. The five *C. maltaromaticum* strains contained the largest secretome (319–375 proteins predicted to harbor a signal peptide), whereas *C. divergens* V41 presented an intermediate number (272 predicted proteins) compared to those predicted in *Carnobacterium* sp. AT7, *Carnobacterium* sp. 17.4, and *C. inhibens* subsp. *gilichinskyi* WN1359 with only 132–155 proteins containing a peptide signal. Compared to other LAB or genera that can share the same habitats, the sizes of secretomes of *C. maltaromaticum* and *C. divergens* are among the largest (**Figure 3**).

Among the COG (cluster of orthologous genes) represented in the secretome, families M (cell wall/membrane/envelop biogenesis), P (inorganic ion transport and metabolism), and R (general functional prediction only) tend to be more represented in *C. maltaromaticum* and *C. divergens* compared to the other species (Supplementary Figure S2). This suggests that extracellular functions are more diversified in *C. maltaromaticum*, and to a lesser extent in *C. divergens*, than in the other *Carnobacterium*. For instance, *C. maltaromaticum* and *C. divergens* exhibit between 22 and 35 genes in COG R, while the three other strains *Carnobacterium* sp. AT7, *Carnobacterium* sp. 17.4, and *C. inhibens* subsp. *gilichinskyi* WN1359, would have less than 10 genes encoding a signal peptide within this subclass.

Further, *C. maltaromaticum* predicted secretomes encompassed a higher number (26–30) of proteins belonging to COG family G (carbohydrate transport and metabolism) than was the case for *C. divergens* V41 (18) and other species (10–15). This is correlated with a higher content in genes encoding PTS transporters, between 62 and 68, in *C. maltaromaticum*.

TABLE 2 | General features of *Carnobacterium* genomes.

Organisms	<i>C. maltaromaticum</i>					<i>C. divergens</i>	<i>C. inhibens</i> subsp. <i>gillichinskyi</i>	<i>Carnobacterium</i> sp.	
	LMA28	DSM20342	ATCC35586	ML.1.97	3.18	V41	WN1359	AT7	17.4
Sequence length (Mbp)	3.65	3.877	3.54	3.33	3.57	2.74	2.35	2.45	2.63
GC content (%)	34.5	34.4	34.5	34.4	34.4	35.3	35.2	35.2	35.2
Number of plasmids	3	ND	ND	ND	ND	ND	5	ND	1
Number of CDS	3,687	3,812	3,448	3,368	3,465	2,633	2,268	2,431	2,584
Number of COG	2,671	2,800	2,636	2,639	2,672	2,089	2,257	1,986	2,155
fCDS	48	12	12	49	8	15	69	12	25
Number of tRNA	59	64	61	37	59	8	75	71	67
Number of 16S-RNA	6	6	ND	ND	ND	ND	8	7	8
Prophage clusters	2*	2 (2)	1 (4)	1	1 (5)	0 (3)	0	0 (3)	0
	(3**)								
Scaffolds	1	5	74	229	160	32	1	69	1
Contigs	1	5	74	229	160	32	1	69	1

* intact and ** region incomplete; ND, not determined.

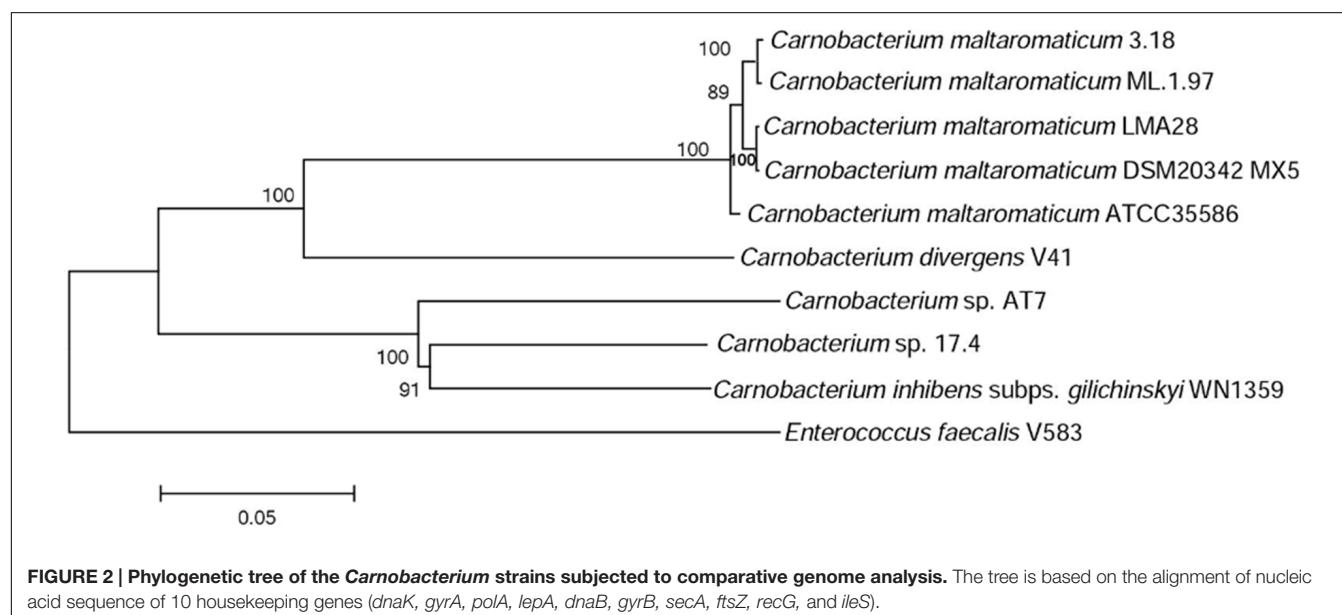


FIGURE 2 | Phylogenetic tree of the *Carnobacterium* strains subjected to comparative genome analysis. The tree is based on the alignment of nucleic acid sequence of 10 housekeeping genes (*dnaK*, *gyrA*, *polA*, *lexA*, *dnaB*, *gyrB*, *secA*, *ftsZ*, *recG*, and *ileS*).

By contrast, the other *Carnobacterium*, including *C. divergens*, would contain less of such genes (between 27 and 43 depending on the strains). *C. maltaromaticum* strains, compared to other *Carnobacterium* species harbor a larger repertoire of PTS transporters. This characteristics is typical of ubiquitous LAB, as these transporters enable the bacteria to exploit a wide range of carbon sources (Douillard and de Vos, 2014).

More strikingly, *C. maltaromaticum* strains and *C. divergens* V41 secretomes encompassed between 116 and 155 proteins unclassified in COG families (class X Supplementary Figure S2), while this category was of minor importance (0–23) in other *Carnobacterium* species. The detailed analysis of those unclassified genes revealed that approximately half of them encode hypothetical proteins or conserved exported proteins of unknown function, and 30% encode secreted proteins associated with the cell wall. These results strongly suggest

that *C. maltaromaticum* and *C. divergens* cell-surface structures differ significantly from those of *Carnobacterium* sp. AT7, *Carnobacterium* sp. 17.4, and *C. inhibens* subsp. *gillichinskyi* WN1359. Therefore, we focused on the comparison of the gene repertoire encoding such surface associated proteins.

Non-covalent Cell-Wall Bound Proteins

A larger set of proteins non-covalently bound to the cell wall was predicted for *C. maltaromaticum* and *C. divergens* V41 than for the other *Carnobacterium*. These proteins contain at least one LysM domain, and one WxL domain or SH3 domain (Table 3).

For all strains, at least one protein with an SH3 domain involved in peptidoglycan binding is predicted to be anchored to the cell wall (Table 3 and Supplementary Table S1), and between four and six LysM proteins. Two of them are conserved in all strains, three are found only in

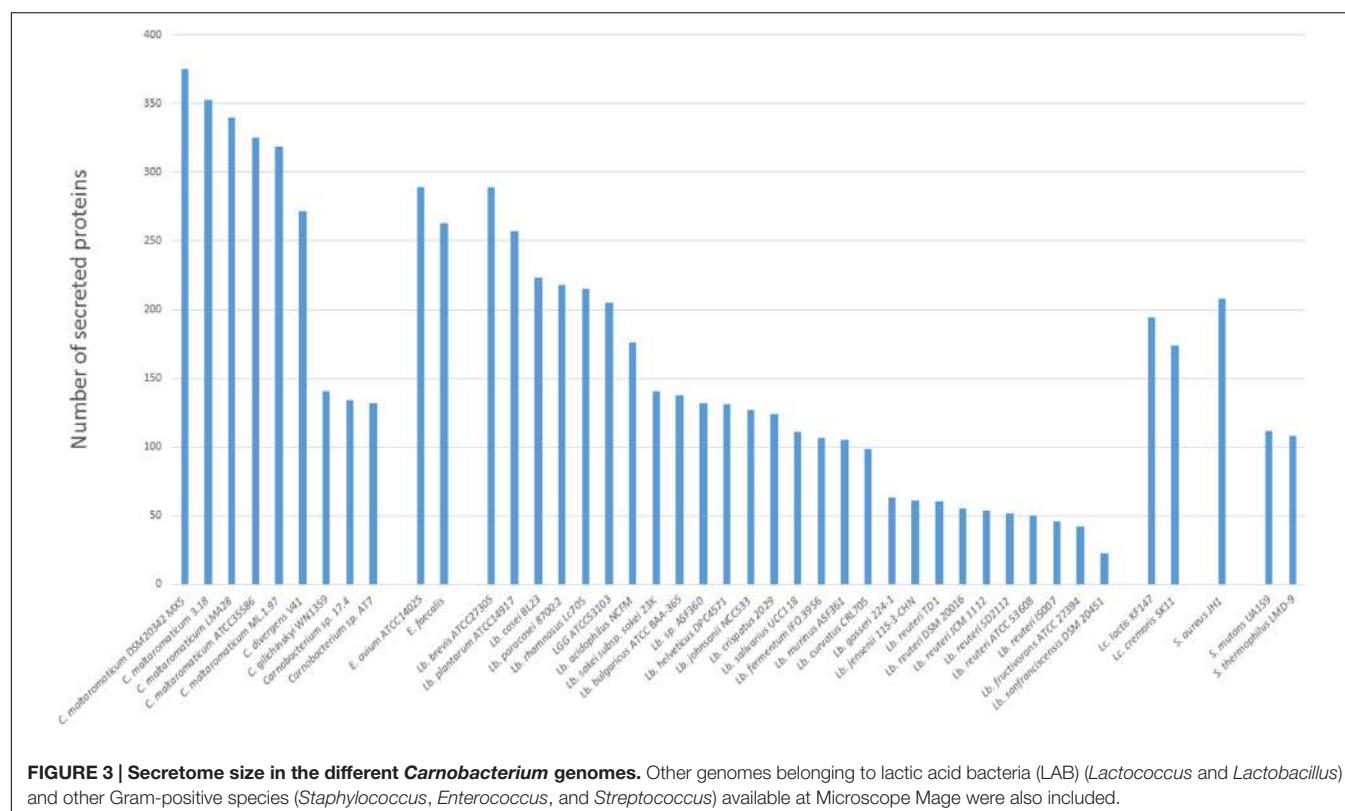


FIGURE 3 | Secretome size in the different *Carnobacterium* genomes. Other genomes belonging to lactic acid bacteria (LAB) (*Lactococcus* and *Lactobacillus*) and other Gram-positive species (*Staphylococcus*, *Enterococcus*, and *Streptococcus*) available at Microscope Mage were also included.

TABLE 3 | Number of sortases and surface proteins.

<i>C. maltaromaticum</i>					<i>C. divergens</i>	<i>C. inhibens</i> subsp. <i>gilichinskyi</i>	<i>Carnobacterium</i> sp.	
LMA28	DSM20342 MX5	ATCC35586	3.18	ML.1.97	V41	WN1359	AT7	17.4
LysM	6	5	5	5	6	4	5	5
WxL	48	49	37	53	46	0	0	0
SH3	3	3	1	2	2	1	1	1
SDP	35	35	28	29	26	0	1	1
Sortase	6	8	6	6	5	0	1	1
Total	98	100	77	95	80	6	7	7

C. maltaromaticum and *C. divergens* V41, and three are variable within *Carnobacterium* sp. strains (Table 3 and Supplementary Table S1).

The signature of the 160–190 aa long WxL domain is characterized by two WxL motifs. WxL-containing proteins are non-covalently anchored proteins associated with the cell wall (Brinster et al., 2007). Strikingly, the number of WxL-containing proteins is comprised of between 37 and 53 and present only in *C. maltaromaticum* and *C. divergens*. Remarkably, no WxL-containing proteins were found to be encoded by the other *Carnobacterium* genomes. Twenty-nine genes encoding WxL proteins are common to the five *C. maltaromaticum* strains and *C. divergens* V41. Overall, 56 genes are variable within the strains (i.e., are absent in at least one strain), and three belong to the *C. maltaromaticum* core genome (Table 3 and Supplementary Table S1). The WxL proteins belong to the cell-surface complex

(Csc) protein family. The Csc protein encoding genes are typically clustered and each cluster contains at least one copy of *cscA*, *cscB*, *cscC*, and *cscD*. Similarly, *C. maltaromaticum* and *C. divergens* contain between 13 and 17 csc clusters. Typically, CscA contains a DUF916 domain with extracellular matrix binding ability (Galloway-Peña et al., 2015) and a C-terminal transmembrane anchor, while CscB and CscC contain WxL domains, and CscD is a small LPXTG protein (Siezen et al., 2006). Similarly, in *C. maltaromaticum* and *C. divergens*, all the WxL encoding genes are either *cscB* or *cscC* homologs, and are localized in the vicinity of at least one WxL encoding gene. These *cscB* and *cscC* homologs can also be clustered with homologs of *cscA* and *cscD*. Strikingly, the clusters can encode a high number of WxL proteins, as exemplified by the cluster BN424_324-BN424_330 which is predicted to encode six WxL proteins.

Whereas *C. maltaromaticum* and *C. divergens* are predicted to produce a high diversity of non-covalently bound proteins, only a small number of such proteins were found in *C. inhibens* subsp. *gilichinskyi* WN1359, *Carnobacterium* sp. AT7, and *Carnobacterium* sp. 17. Among those, most are predicted as LysM- and SH3-containing proteins and no WxL proteins were detected (**Table 3** and Supplementary Table S1).

Covalently Anchored Proteins

Sortase-dependent proteins (SDP) are covalently anchored to the cell wall, and possess an LPxTG like motif at their C-terminal end. SDP nomenclature refers to proteins attached to the peptidoglycan by the sortase family of transpeptidases (Schneewind and Missiakas, 2014). Such SDP were found in *C. maltaromaticum* strains and *C. divergens* V41 whereas only one LPxTG domain protein was detected in *Carnobacterium* sp. 17.4 and *Carnobacterium* sp. AT7, and none in *C. inhibens* subsp. *gilichinskyi* WN1359 (**Table 3** and Supplementary Table S1).

Sortases decorate the surfaces of Gram-positive bacteria with diverse proteins that enable microbes to interact with their environment (Comfort and Clubb, 2004; Maresso and Schneewind, 2008). The five *C. maltaromaticum* strains and *C. divergens* V41 possess many putative sortase A and B genes, while *Carnobacterium* sp. AT7 and *Carnobacterium* sp. 17.4 harbor only one and *C. inhibens* subsp. *gilichinskyi* WN1359 possesses a pseudogene that might encode a remnant protein with similarities with sortases (**Table 3** and Supplementary Table S1).

Depending on the strains, the *C. maltaromaticum* genomes are predicted to encode 26–35 SDP, and 26 putative SDP could be predicted in *C. divergens*. Among those, 13 belong to the *C. maltaromaticum* core genome, including seven also conserved in *C. divergens*. In addition, 24 *C. maltaromaticum* SDP-encoding genes are strain specific or shared by only some of the strains. The differences between strains resulted either from the absence of homologs or the presence of 373 predicted pseudogenes. The predicted sortase-encoding gene in the genome of *Carnobacterium* sp. 17.4 is located next to a collagen-binding surface protein encoding gene (ABHHv1_120049, Supplementary Table S1).

Functions of the Surface Proteins

As *C. maltaromaticum* and, to a lesser extent, *C. divergens* presented a large panel of surface-exposed proteins compared to other *Carnobacterium* species, we searched for the putative functions of the *C. maltaromaticum* and *C. divergens* species specific proteins to illuminate the potential benefits they might provide to these two species.

Enzymes

Homologs of multidomain proteins predicted as nucleotidases/metallophosphatases as well as PrtB homologs are among the proteins conserved in *C. maltaromaticum* and *C. divergens* (**Figure 4** and Supplementary Table S1). Extracellular 5'-nucleotidase domains that catalyze dephosphorylation of exogenous adenine 5'-nucleotides to adenosine and phosphate (Bengis-Garber and Kushner, 1982) are reported as providing a

key function for phosphorous regeneration in aquatic ecosystems (Ammerman and Azam, 1985).

In the dairy LAB *Lactobacillus delbrueckii* subsp. *bulgaricus*, PrtB, a cell envelope-associated protease (CEP) has been shown to degrade casein into peptides (Siezen, 1999). Peptides and aas are subsequently internalized and peptides are further hydrolyzed by intracellular peptidases into small peptides and free aa (Savijoki et al., 2006). The analysis of the genomes of *C. maltaromaticum* and *C. divergens* revealed the conserved presence of oligopeptides transporter systems OppABCDF and DtpT, as well as intracellular peptidases (Supplementary Table S1). Interestingly, homologs of the Opp system and of intracellular peptidases were also found in *Carnobacterium* sp. 17.4, *Carnobacterium* sp. AT7, and *C. inhibens* subsp. *gilichinskyi* WN1359 but no homologs of CEP and no DtpT. In the dairy environment, CEP are believed to confer a selective advantage by allowing bacteria to exploit the aa from milk casein (Kunji et al., 1996; Christensen et al., 1999; Doeven et al., 2005; Savijoki et al., 2006). All these data strongly suggest that among *Carnobacterium* only *C. maltaromaticum* and *C. divergens* are indeed able to exploit aa from the proteins present in their environments. The presence of PrtB would be a selective advantage for these two species in protein-rich environments such as food.

Among the SDP proteins present in several *C. maltaromaticum* strains we noticed a CDS predicted as harboring a glycoside hydrolase activity (BN424_641, **Figure 4** and Supplementary Table S1). This multidomain protein also contain a fibronectin type III-like module of unknown function, which is usually associated with glycoside hydrolase domains (Alahuhta et al., 2010). Such enzymatic activity could allow *C. maltaromaticum* to degrade extracellular carbohydrate polymers. Interestingly, this protein is also predicted to contain a mucin-binding domain. Mucin is the major component of the intestinal mucus. It is tempting to speculate that the putative mucin-binding glycoside hydrolase of *C. maltaromaticum* would hydrolyze mucin glycan moieties as has been described for some gut bacteria (Tailford et al., 2015).

Nutrient Uptake: Heme Compounds

Two SDP, predicted as heme-binding proteins in *C. maltaromaticum* and *C. divergens*, are homologs of IsdA and IsdC (**Figure 5** and Supplementary Table S1). The Isd system in *Staphylococcus aureus* enables to utilize different sources of heme: free heme, heme bound to free hemoglobin, and heme bound to hemoglobin interacting with haptoglobin. The Isd ABC-transport system in *S. aureus* is encoded by *isdABCDEFGH*; IsdH, IsdB, and IsdA acting as cell wall anchored receptor proteins: IsdH is the primary haptoglobin-hemoglobin receptor, IsdB the primary hemoglobin receptor, and IsdA can bind free heme or accept heme from IsdB or IsdH. Heme is subsequently transferred from IsdA to IsdC, another cell wall protein and then to a membrane associated transporter, formed by IsdD, IsdE, and IsdF. After internalization, heme is taken up by the heme degrading monooxygenase IsdG that releases the resulting iron in the cytoplasm (Choby and Skaar, 2016). At first glance, among the possible protein candidates able to bind heme from the environment, *C. maltaromaticum* and *C. divergens* would

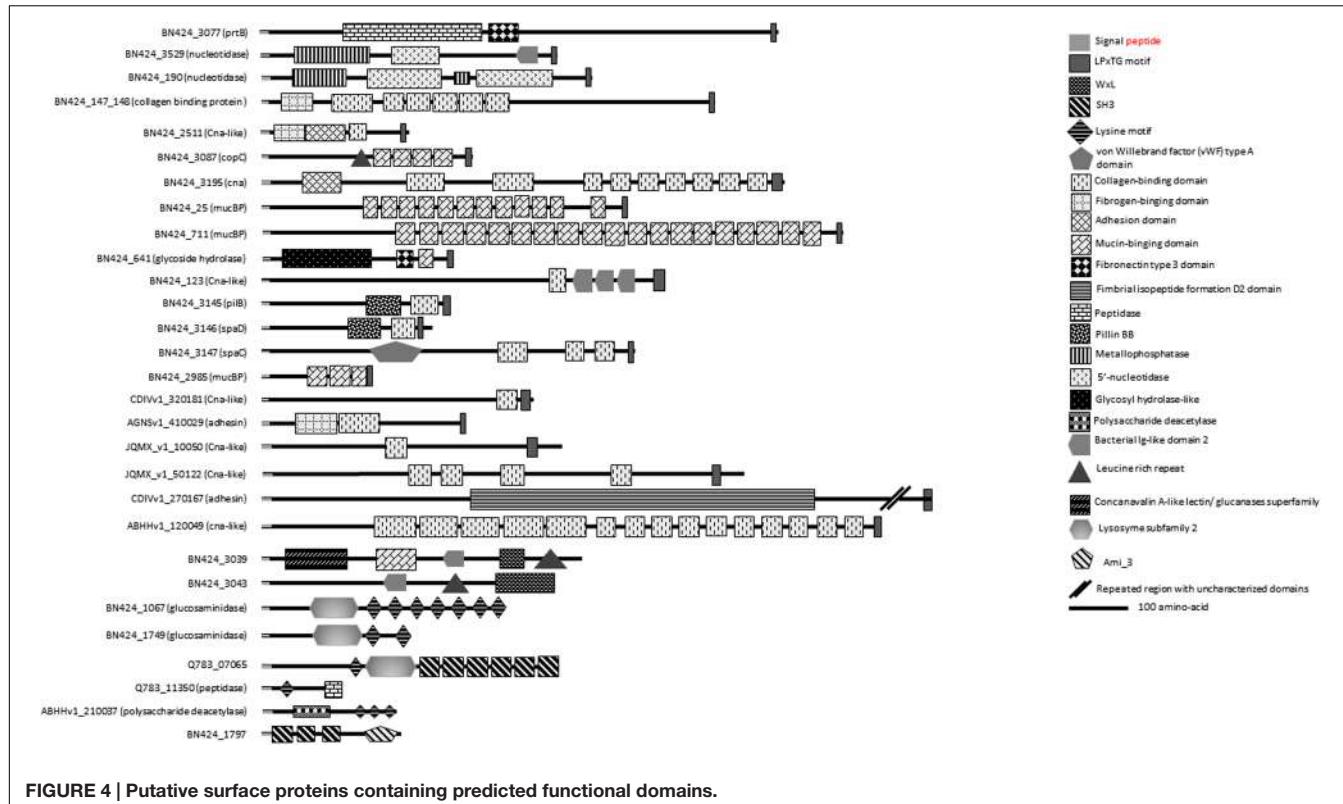


FIGURE 4 | Putative surface proteins containing predicted functional domains.

only have IsdA as no ortholog of IsdB and IsdH were found. This would suggest that these two *Carnobacterium* species would primarily be able to use free heme as iron source and not heme bound to proteins. However, the IsdA homolog in *C. maltaromaticum* is predicted to contain four NEAT domains while IsdA from *S. aureus* contains only two. NEAT domains bind heme compounds or proteins containing heme compound (Gaudin et al., 2011; Balderas et al., 2012). It could therefore be speculated that the presence of two additional NEAT domains in the *C. maltaromaticum* IsdA homolog might compensate the absence of IsdB.

Carnobacterium divergens and *C. maltaromaticum* do not exhibit any IsdD homolog. However, we found two *fhuC* homologs which encode an ATP-binding component of ABC transporter as well. These *FhuC* homologs could play the equivalent role of IsdD and thus build a functional ABC transporter with the permease encoded by the IsdF homolog. In *C. divergens* V41, one *fhuC* is localized in the vicinity of the putative *isd* genes, while in the *C. maltaromaticum* strains, this homolog is localized elsewhere in the genome. In addition, all the analyzed genomes of *Carnobacterium* contain an *isdG* homolog and would be able to release iron after heme import in the cytoplasm. The *isdA* homolog of *C. maltaromaticum* ML1.97 and 3.18 appear as a pseudogene, indicating that this system is not fully functional in these two strains. Overall, these analyses suggest *C. divergens* V41 and several strains if not all of *C. maltaromaticum* would be able to use extracellular heme, for respiration (see below) and/or as an iron source.

Microbial Adhesion

The previous analysis of the genome of *C. maltaromaticum* LMA28 reported the presence of putative cell-surface adhesins (Rahman et al., 2014b). Comparison of *Carnobacterium* genomes revealed that *C. maltaromaticum* strains and *C. divergens* V41 may produce several surface proteins predicted to contain domains reminiscent of adhesion function: collagen-binding, mucBP, Leucine-Rich Repeat (LRR).

Among those, 10 LPxTG proteins are predicted to have a collagen-binding domain (Supplementary Table S1 and Figure 4). Functional domains are annotated collagen-binding protein, Cna, or bacterial adhesin. Collagen-binding proteins found in *S. aureus* (Elasri et al., 2002) and *Listeria monocytogenes* (Bierne and Cossart, 2007) are suggested to participate in the infection process. However, proteins putatively involved in adhesion to collagen and mucin have also been reported to be important for the probiotic properties of LAB as shown in *Lactobacillus plantarum* WCFS1 (Boekhorst et al., 2006b). It is therefore difficult to predict if the binding capacity of such proteins in *Carnobacterium* can be considered as beneficial or not.

Three LPxTG proteins (Figure 4 and Supplementary Table S2) contain a MucBP (mucin-binding protein) domain. MucBP domains allow adhesion to mucus material (Lukić et al., 2012). *Lactobacillales* and *Listeria* species can possess between 1 and 14 MucBP-containing proteins (Boekhorst et al., 2006a; Bierne and Cossart, 2007).

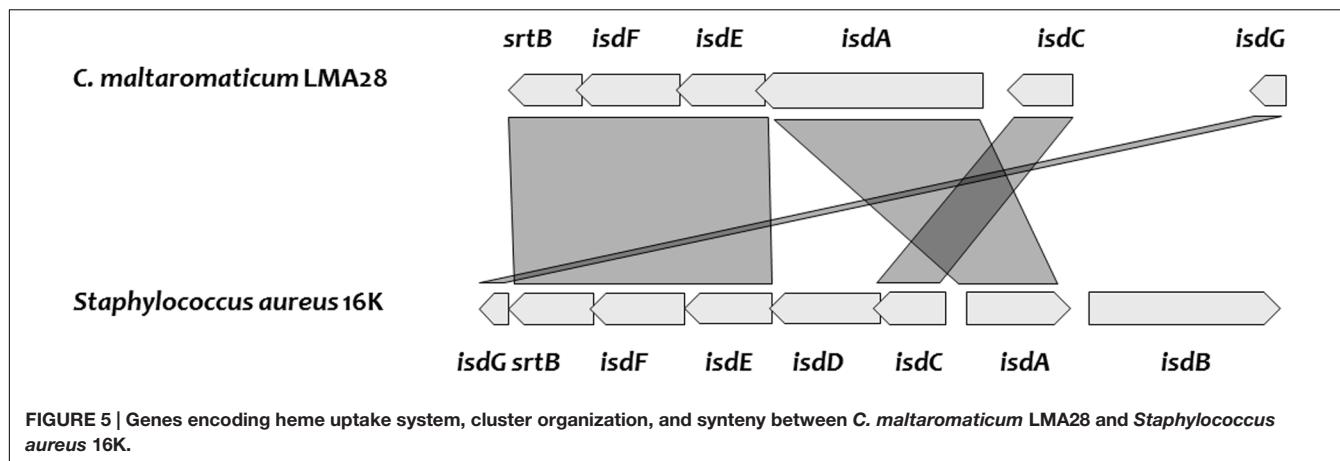


FIGURE 5 | Genes encoding heme uptake system, cluster organization, and synteny between *C. maltaromaticum* LMA28 and *Staphylococcus aureus* 16K.

In *Enterococcus faecium*, the WxL protein SwpA and the protein DufA, which contains a DUF916 domain and is encoded by a gene that belongs to a *csc* cluster, are collagen and fibronectin-binding proteins (Galloway-Peña et al., 2015). Similarly, it can be hypothesized that at least some WxL and DUF916 likely encoded by *C. divergens* and *C. maltaromaticum* might exhibit similar adhesion properties. In addition, two LRR domains were found in some strains of *C. maltaromaticum* holding a WxL anchorage domain (Figure 4 and Supplementary Table S2). These domains are involved in protein–protein interactions. They are described to be associated with domains exhibiting an Ig-like (immunoglobulin) fold, the function of which is thought to facilitate the presentation of the adjacent LRR domain (Bierne and Cossart, 2007). Accordingly, near the LRR domain of the putative *C. maltaromaticum* LMA28 surface protein BN424_3043, an Ig-like domain was found at the C-terminal end of the LRR region (Figure 4 and Supplementary Table S2). In *L. monocytogenes*, internalins associated to virulence are LRR-containing proteins. LRR-containing proteins are rather uncommon in *Lactobacillus*.

The presence of such putative adhesins suggests that *C. maltaromaticum* and *C. divergens* exhibit the ability to adhere to intestinal mucosa and extracellular matrices of animals. Overall, 19 putative adhesins were predicted from the genomes of *C. maltaromaticum* and *C. divergens* and absent from the other *Carnobacterium* species. All these proteins are SDP except two LRR-containing proteins which exhibit a WxL binding domain. Of these 19 proteins, two are conserved in all *C. maltaromaticum* strains and absent in other *Carnobacterium*: a putative collagen-binding SDP and a mucin-binding protein.

Genes encoding the pili proteins previously described for *C. maltaromaticum* LMA28 (Rahman et al., 2014b), were found only in one other *C. maltaromaticum* strain (DSM20342 MX5). Both strains belong to clonal complex CC1, which is suspected to be a lineage well-adapted to the dairy environment. Pili were described as surface components promoting adhesion to dairy matrix in the probiotic strain *L. rhamnosus* GG and thereby they might contribute to confer an advantage in dairy products (Burgain et al., 2014). In general in Gram-positive bacteria, two or three genes encoding the pilus subunits are organized into

an operon, along with at least one sortase gene (Mandlik et al., 2008; Proft and Baker, 2009). The closest homologs of such *C. maltaromaticum* genes were found in *E. faecalis*, with also a highly similar genetic organization (Figure 6). However, pilin gene organization is different between *C. maltaromaticum* and *L. rhamnosus* GG suggesting that the genetic structures of pili loci in *C. maltaromaticum*, *E. faecalis* V583, and *L. rhamnosus* GG are the result of different gene rearrangements.

Overall, these comparative genomic analyses suggest that all *C. maltaromaticum* strains and *C. divergens* might have adhesive properties and that strains might exhibit differences in this regard.

Further, these analyses demonstrated striking differences between the group *C. maltaromaticum/C. divergens* and the other *Carnobacterium* spp. The secretome and more specifically the cell-surface proteome of *C. maltaromaticum* and *C. divergens* are large, the one of *C. maltaromaticum* being the largest described for LAB as previously noticed by Sun et al. (2015). Conversely the secretome of *C. inhibens* subsp. *gillichinskyi* WN1359, *Carnobacterium* sp. AT7, and *Carnobacterium* sp. 17.4 is the smallest among LAB. The detailed analysis of the functions provided by such large secretome supports a prediction that the cell-surface proteome would confer the ability to hydrolyze and to adhere to biomacromolecules as well as to capture biomolecules (heme compounds). The cell surface of *C. maltaromaticum* and *C. divergens* might closely interact with animal environments and use these nutrient-rich substrates. In addition, *C. maltaromaticum* would likely exhibit a larger repertoire of hydrolytic enzymes and adhesins that may enable *C. maltaromaticum* to adapt to multiple habitats. By contrast, the other *Carnobacterium* lack such a cell-surface proteome and would therefore be expected to be less able to colonize animal-linked habitats. The size of the secretome is highly variable in LAB. Interestingly, all known specialized LAB *Streptococcus thermophilus*, *L. iners*, *L. reuteri*, *L. sanfranciscensis*, and *L. fructivorans* are characterized by a small secretome, suggesting that colonizing one specific niche, such as dairy, vagina, sourdough, or the GI tract, respectively, does not require a large secretome. Consistently, the known generalist LAB such as *L. rhamnosus*, *L. plantarum* (Martino et al., 2016), and some

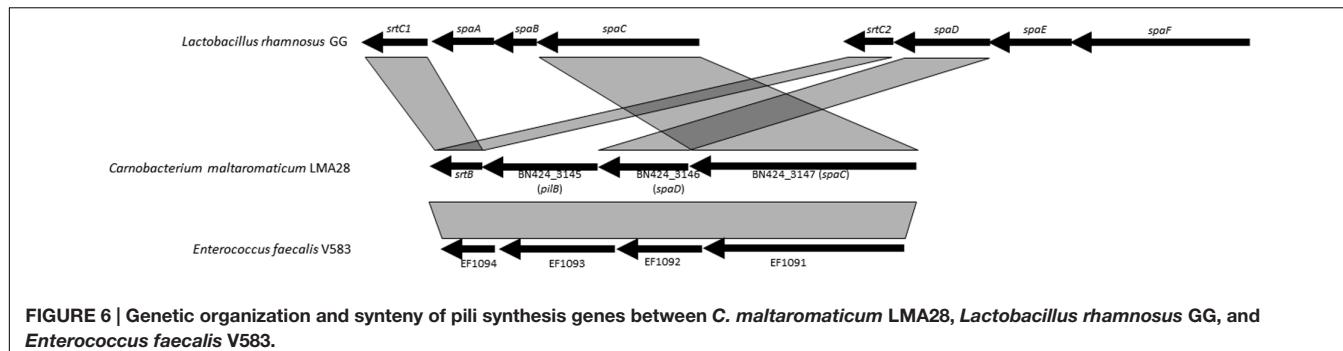


FIGURE 6 | Genetic organization and synteny of pili synthesis genes between *C. maltaromaticum* LMA28, *Lactobacillus rhamnosus* GG, and *Enterococcus faecalis* V583.

Enterococcus spp. exhibit a large secretome. We propose that there is an intimate relationship between the secretome size and the ability of LAB to colonize diverse habitats. According to this hypothesis, the bigger a secretome, the higher the capacity to colonize multiple environments.

Respiration

Carnobacterium maltaromaticum is unable to synthesize heme and exhibits better growth efficiency in the presence of hematin, suggesting that heme is used by *C. maltaromaticum* to respire oxygen. Consistent with this hypothesis, *C. maltaromaticum* has been reported to produce cytochrome b and d types when grown aerobically with hematin (Meisel et al., 1994). The gene repertoire of *C. maltaromaticum* suggests the ability to produce a functional respiratory chain. Indeed, the electron donor-encoding genes, *noxB* and *ndh*, and those required for the synthesis of the electron shuttle menaquinone (*yhdB*, *menE*, *menB*, *menD*, *menF*, *ispA*, *ispB*, and *menA*) were present. Further, *cydABCD*, encoding the heme-dependent cytochrome quinol oxidase that performs the final electron transfer to the acceptor oxygen (Lechardeur et al., 2011), were also found. These genes were present in all *C. maltaromaticum* strains and in *C. divergens*, except in *C. maltaromaticum* ML.1.97 whose *menF* appears to be a pseudogene. Therefore, this last strain might require the presence of quinone in the environment to respire, as reported for *Streptococcus agalactiae* (Rezaïki et al., 2008). It seems therefore that the components of the cell wall proteome involved in extracellular heme utilization may also contribute to oxygen respiration in *C. maltaromaticum* and *C. divergens*.

By contrast, only *menA*, *noxA/noxB*, and *ispB* were found in *Carnobacterium* sp. AT7, *Carnobacterium* sp. 17.4, and *Carnobacterium inhibens* subsp. *gilichinskyi* WN1359 strongly suggesting that these three bacteria do not possess any functional respiratory chain.

Adaptation to the GI Tract

Bacteria have to deal with several stresses in order to survive in the GI tract, including the immune system and bile. All strains of *C. maltaromaticum* and *C. divergens* contain genes conferring resistance to some components of the immune system. Indeed, homologs were found for *mprF*, *dltABCD*, *asnH/asnB*, *oatA*, and *pgdA* (Supplementary Table S2). The genes *mprF* and *dltABCD* are involved in phospholipid lysinylation

and teichoic acid D-alanylation, respectively, and thus confer resistance to antimicrobial peptides of the innate immune system by protecting the cell wall. The genes *oatA* and *pgdA* confer to peptidoglycan a high resistance to lysozyme, another component of the innate immune system, by introducing O-acetylation and N-deacetylation, respectively. Importantly, no *mprF*, *dltABCD*, *asnH/asnB*, *oatA*, and *pgdA* homologs were found in the genomes of *Carnobacterium* sp. AT7, *Carnobacterium* sp. 17.4, and *Carnobacterium inhibens* subsp. *gilichinskyi* WN1359.

Bile salt hydrolase encoding genes were conserved in all strains of *C. maltaromaticum* while none were found in other Carnobacteria including *C. divergens* V41 (Supplementary Table S2).

Carnobacterium maltaromaticum and *C. divergens* are thus highly contrasting from the three other *Carnobacterium* taxons by their content of genes described in *Lactobacillus* as key factor for survival in the GI tract of animals (Kleerebezem et al., 2010). Indeed, while *C. maltaromaticum* and *C. divergens* possess genes putatively conferring resistance against the innate immune system, almost none of these homologs were found in *Carnobacterium* sp. AT7, *Carnobacterium* sp. 17.4, and *C. inhibens* subsp. *gilichinskyi* WN1359. This might explain why the only two species identified in feces samples are *C. maltaromaticum* and *C. divergens*. However, surprisingly, genes allowing resistance to bile were only found in *C. maltaromaticum* and not in *C. divergens*. Yet the ability to hydrolyze bile is described as a key factor for colonization of the gut (Kleerebezem et al., 2010; Seedorf et al., 2014). Indeed *C. maltaromaticum* LMA28, a cheese isolate that possesses such genes, is able to survive during the gastrointestinal transit in a mouse model (Rahman et al., 2014b; Sun et al., 2015). However, such ability has not yet been tested in *C. divergens* V41. Whether these two species differ in their ability to deal with bile and the ecological consequences it has on GI tract survival deserves further investigation.

AUTHOR CONTRIBUTIONS

Performed 16S meta-barcoding: BT and GD. Performed whole genome sequencing and assembly: JL, MH, and SS. Comparative genome analyses: CI, CC-G, FB, A-MR-J, MZ, and BR. Wrote the manuscript: CI, CC-G, BT, MZ, BR, JL, CM, FB, and A-MR-J. Coordinated the study: FB.

ACKNOWLEDGMENTS

The LABGeM (CEA/IG/Genoscope & CNRS UMR8030) and the France Génomique National infrastructure (funded as part of Investissement d'avenir program managed by Agence Nationale pour la Recherche, contract ANR-10-INBS-09) are acknowledged for support within the MicroScope annotation platform. They are also thankful to Myriam Michelle, Sylvie Wolff, Arnaud Khemisti, and Camille Collin for their technical support during this study. The authors would like to thank

the reviewers for their helpful and constructive comments that greatly contributed to improving the final version of the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.00357/full#supplementary-material>

REFERENCES

- Alahuhta, M., Xu, Q., Brunecky, R., Adney, W. S., Ding, S.-Y., Himmel, M. E., et al. (2010). Structure of a fibronectin type III-like module from *Clostridium thermocellum*. *Acta Cryst. F* 66, 878–880. doi: 10.1107/S1744309110022529
- Ammerman, J. W., and Azam, F. (1985). Bacterial 5'-nucleotidase in aquatic ecosystems: a novel mechanism of phosphorus regeneration. *Science* 227, 1338–1340. doi: 10.1126/science.227.4692.1338
- Balderas, M. A., Nobles, C. L., Honsa, E. S., Alicki, E. R., and Maresso, A. W. (2012). Hal is a *Bacillus anthracis* heme acquisition protein. *J. Bacteriol.* 194, 5513–5521. doi: 10.1128/JB.00685-12
- Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712. doi: 10.1126/science.1138140
- Bengis-Garber, C., and Kushner, D. J. (1982). Role of membrane-bound 5'-nucleotidase in nucleotide uptake by the moderate halophile *Vibrio costicola*. *J. Bacteriol.* 149, 808–815.
- Bierne, H., and Cossart, P. (2007). Listeria monocytogenes surface proteins: from genome predictions to function. *Microbiol. Mol. Biol. Rev.* 71, 377–397. doi: 10.1128/MMBR.00039-06
- Boekhorst, J., Helmer, Q., Kleerebezem, M., and Siezen, R. J. (2006a). Comparative analysis of proteins with a mucus-binding domain found exclusively in lactic acid bacteria. *Microbiology* 152, 273–280.
- Boekhorst, J., Wels, M., Kleerebezem, M., and Siezen, R. J. (2006b). The predicted secretome of *Lactobacillus plantarum* WCFS1 sheds light on interactions with its environment. *Microbiology* 152, 3175–3183.
- Brinster, S., Furlan, S., and Serror, P. (2007). C-terminal WxL domain mediates cell wall binding in *Enterococcus faecalis* and other gram-positive bacteria. *J. Bacteriol.* 189, 1244–1253. doi: 10.1128/JB.00773-06
- Burgain, J., Scher, J., Lebeer, S., Vanderleyden, J., Cailliez-Grimal, C., Corgneau, M., et al. (2014). Significance of bacterial surface molecules interactions with milk proteins to enhance microencapsulation of *Lactobacillus rhamnosus* GG. *Food Hydrocoll.* 41, 60–70. doi: 10.1016/j.foodhyd.2014.03.029
- Cailliez-Grimal, C., Afzal, M. I., and Revol-Junelles, A.-M. (2014). “Carnobacterium,” in *Encyclopedia of Food Microbiology*, eds C. A. Batt and M. L. Tortorello (Cambridge, MA: Academic Press), 379–383. doi: 10.1016/B978-0-12-384730-0.00381-5
- Cailliez-Grimal, C., Chaillou, S., Anba-Mondoloni, J., Loux, V., Afzal, M. I., Rahman, A., et al. (2013). Complete chromosome sequence of *Carnobacterium maltaromaticum* LMA 28. *Genome Announc.* 1:e00115-12. doi: 10.1128/genomeA.00115-12
- Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 42, D459–D471. doi: 10.1093/nar/gkt1103
- Chaillou, S., Chaulot-Talmon, A., Caekebeke, H., Cardinal, M., Christeans, S., Denis, C., et al. (2015). Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. *ISME J.* 9, 1105–1118. doi: 10.1038/ismej.2014.202
- Chobay, J. E., and Skaar, E. P. (2016). Heme synthesis and acquisition in bacterial pathogens. *J. Mol. Biol.* 428, 3408–3428. doi: 10.1016/j.jmb.2016.03.018
- Christensen, J. E., Dudley, E. G., Pederson, J. A., and Steele, J. L. (1999). Peptidases and amino acid catabolism in lactic acid bacteria. *Antonie Van Leeuwenhoek* 76, 217–246. doi: 10.1023/A:1002001919720
- Comfort, D., and Clubb, R. T. (2004). A comparative genome analysis identifies distinct sorting pathways in Gram-positive bacteria. *Infect. Immun.* 72, 2710–2722. doi: 10.1128/IAI.72.5.2710-2722.2004
- Delorme, C., Bartholini, C., Bolotine, A., Ehrlich, S. D., and Renault, P. (2010). Emergence of a cell wall protease in the *Streptococcus thermophilus* population. *Appl. Environ. Microbiol.* 76, 451–460. doi: 10.1128/AEM.01018-09
- Doeven, M. K., Kok, J., and Poolman, B. (2005). Specificity and selectivity determinants of peptide transport in *Lactococcus lactis* and other microorganisms: peptide transport in *Lactococcus lactis*. *Mol. Microbiol.* 57, 640–649. doi: 10.1111/j.1365-2958.2005.04698.x
- Douglas, G. L., and Klaenhammer, T. R. (2010). Genomic evolution of domesticated microorganisms. *Annu. Rev. Food Sci. Technol.* 1, 397–414. doi: 10.1146/annurev.food.102308.124134
- Douillard, F. P., and de Vos, W. M. (2014). Functional genomics of lactic acid bacteria: from food to health. *Microb. Cell Fact.* 13:S8. doi: 10.1186/1475-2859-13-S1-S8
- Douillard, F. P., Ribbera, A., Kant, R., Pietila, T. E., Jarvinen, H. M., Messing, M., et al. (2013). Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG. *PLoS Genet.* 9:e1003683. doi: 10.1371/journal.pgen.1003683
- Duan, S., Hu, X., Li, M., Miao, J., Du, J., and Wu, R. (2016). Composition and metabolic activities of the bacterial community in shrimp sauce at the flavor-forming stage of fermentation as revealed by metatranscriptome and 16S rRNA gene sequencing. *J. Agric. Food Chem.* 64, 2591–2603. doi: 10.1021/acs.jafc.5b05826
- Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimer detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381
- Elasri, M. O., Thomas, J. R., Skinner, R. A., Blevins, J. S., Beenken, K. E., Nelson, C. L., et al. (2002). *Staphylococcus aureus* collagen adhesin contributes to the pathogenesis of osteomyelitis. *Bone* 30, 275–280. doi: 10.1016/S8756-3282(01)00632-9
- Fougy, L., Desmonts, M.-H., Coeuillet, G., Fassel, C., Hamon, E., Hézard, B., et al. (2016). Reducing salt in raw pork sausages increases spoilage and correlates with reduced bacterial diversity. *Appl. Environ. Microbiol.* 82, 3928–3939. doi: 10.1128/AEM.00323-16
- Frese, S. A., Benson, A. K., Tannock, G. W., Loach, D. M., Kim, J., Zhang, M., et al. (2011). The evolution of host specialization in the vertebrate gut symbiont *Lactobacillus reuteri*. *PLOS Genet.* 7:e1001314. doi: 10.1371/journal.pgen.1001314
- Galloway-Peña, J. R., Liang, X., Singh, K. V., Yadav, P., Chang, C., Rosa, S. L. L., et al. (2015). The identification and functional characterization of WxL proteins from *Enterococcus faecium* reveal surface proteins involved in extracellular matrix interactions. *J. Bacteriol.* 197, 882–892. doi: 10.1128/JB.02288-14
- Gaudin, C. F. M., Grigg, J. C., Arrieta, A. L., and Murphy, M. E. P. (2011). Unique heme-iron coordination by the hemoglobin receptor IsdB of *Staphylococcus aureus*. *Biochemistry* 50, 5443–5452. doi: 10.1021/bi200369p
- Hiu, S. F., Holt, R. A., Sriranganathan, N., Seidler, R. J., and Fryer, J. L. (1984). *Lactobacillus piscicola*, a new species from salmonid fish. *Int. J. Syst. Bacteriol.* 34, 393–400. doi: 10.1099/00207713-34-4-393
- Iskandar, C. F., Cailliez-Grimal, C., Rahman, A., Rondags, E., Remenant, B., Zagorec, M., et al. (2016). Genes associated to lactose metabolism illustrate the high diversity of *Carnobacterium maltaromaticum*. *Food Microbiol.* 58, 79–86. doi: 10.1016/j.fm.2016.03.008

- Jääskeläinen, E., Hultman, J., Parshintsev, J., Riekkola, M.-L., and Björkroth, J. (2016). Development of spoilage bacterial community and volatile compounds in chilled beef under vacuum or high oxygen atmospheres. *Int. J. Food Microbiol.* 223, 25–32. doi: 10.1016/j.ijfoodmicro.2016.01.022
- Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27
- Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi: 10.1093/nar/gkt1076
- Kleerebezem, M., Hols, P., Bernard, E., Rolain, T., Zhou, M., Siezen, R. J., et al. (2010). The extracellular biology of the *Lactobacilli*. *FEMS Microbiol. Rev.* 34, 199–230. doi: 10.1111/j.1574-6976.2009.00208.x
- Kunji, E. R., Mierau, I., Hagting, A., Poolman, B., and Konings, W. N. (1996). The proteolytic systems of lactic acid bacteria. *Antonie Van Leeuwenhoek* 70, 187–221. doi: 10.1007/BF00395933
- Lauro, F. M., Chastain, R. A., Blankenship, L. E., Yayanos, A. A., and Bartlett, D. H. (2007). The unique 16S rRNA genes of piezophiles reflect both phylogeny and adaptation. *Appl. Environ. Microbiol.* 73, 838–845. doi: 10.1128/AEM.01726-06
- Laursen, B. G., Bay, L., Cleenwerck, I., Vancanneyt, M., Swings, J., Dalgaard, P., et al. (2005). *Carnobacterium divergens* and *Carnobacterium maltaromaticum* as spoilers or protective cultures in meat and seafood: phenotypic and genotypic characterization. *Syst. Appl. Microbiol.* 28, 151–164. doi: 10.1016/j.syapm.2004.12.001
- Lechardeur, D., Cesselin, B., Fernandez, A., Lamberet, G., Garrigues, C., Pedersen, M., et al. (2011). Using heme as an energy boost for lactic acid bacteria. *Curr. Opin. Biotechnol.* 22, 143–149. doi: 10.1016/j.copbio.2010.12.001
- Leisner, J. J., Hansen, M. A., Larsen, M. H., Hansen, L., Ingmer, H., and Sørensen, S. J. (2012). The genome sequence of the lactic acid bacterium, *Carnobacterium maltaromaticum* ATCC 35586 encodes potential virulence factors. *Int. J. Food Microbiol.* 152, 107–115. doi: 10.1016/j.ijfoodmicro.2011.05.012
- Leisner, J. J., Laursen, B. G., Prévost, H., Drider, D., and Dalgaard, P. (2007). *Carnobacterium*: positive and negative effects in the environment and in foods. *FEMS Microbiol. Rev.* 31, 592–613. doi: 10.1111/j.1574-6976.2007.00080.x
- Leonard, M. T., Panayotova, N., Farmerie, W. G., Triplett, E. W., and Nicholson, W. L. (2013). Complete genome sequence of *Carnobacterium gilichinskyi* strain WN1359T (DSM 27470T). *Genome Announc.* 1:e00985-13. doi: 10.1128/genomeA.00985-13
- Lorca, G. L., Barabote, R. D., Zlotopolski, V., Tran, C., Winnen, B., Hvorup, R. N., et al. (2007). Transport capabilities of eleven gram-positive bacteria: comparative genomic analyses. *Biochim. Biophys. Acta* 1768, 1342–1366. doi: 10.1016/j.bbamem.2007.02.007
- Lukić, J., Strahinić, I., Jovičić, B., Filipić, B., Topisirović, L., Kojić, M., et al. (2012). Different roles for lactococcal aggregation factor and mucin binding protein in adhesion to gastrointestinal mucosa. *Appl. Environ. Microbiol.* 78, 7993–8000. doi: 10.1128/AEM.02141-12
- Macklaim, J. M., Gloor, G. B., Anukam, K. C., Cribby, S., and Reid, G. (2011). At the crossroads of vaginal health and disease, the genome sequence of *Lactobacillus iners* AB-1. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4688–4695. doi: 10.1073/pnas.1000086107
- Mandlik, A., Das, A., and Ton-That, H. (2008). The molecular switch that activates the cell wall anchoring step of pilus assembly in gram-positive bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 105, 14147–14152. doi: 10.1073/pnas.0806350105
- Maresco, A. W., and Schneewind, O. (2008). Sortase as a target of anti-infective therapy. *Pharmacol. Rev.* 60, 128–141. doi: 10.1124/pr.107.07110
- Martino, M. E., Bayjanov, J. R., Caffrey, B. E., Wels, M., Joncour, P., Hughes, S., et al. (2016). Nomadic lifestyle of *Lactobacillus plantarum* revealed by comparative genomics of 54 strains isolated from different habitats. *Environ. Microbiol.* 18, 4974–4989. doi: 10.1111/1462-2920.13455
- Meisel, J., Wolf, G., and Hammes, W. P. (1994). Heme-dependent cytochrome formation in *Lactobacillus maltaromaticus*. *Syst. Appl. Microbiol.* 17, 20–23. doi: 10.1016/S0723-2020(11)80026-3
- Mendes-Soares, H., Suzuki, H., Hickey, R. J., and Forney, L. J. (2014). Comparative functional genomics of *Lactobacillus* spp. reveals possible mechanisms for specialization of vaginal *Lactobacilli* to their environment. *J. Bacteriol.* 196, 1458–1470. doi: 10.1128/JB.01439-13
- Miller, A., Morgan, M. E., and Libbey, L. M. (1974). *Lactobacillus maltaromaticus*, a new species producing a malty aroma. *Int. J. Syst. Bacteriol.* 24, 346–354. doi: 10.1099/00207713-24-3-346
- Millière, J. B., Michel, M., Mathieu, F., and Lefebvre, G. (1994). Presence of *Carnobacterium* spp. in French surface mould-ripened soft-cheese. *J. Appl. Bacteriol.* 76, 264–269. doi: 10.1111/j.1365-2672.1994.tb01626.x
- Nicholson, W. L., Zhahnina, K., de Oliveira, R. R., and Triplett, E. W. (2015). Proposal to rename *Carnobacterium inhibens* as *Carnobacterium inhibens* subsp *inhibens* subsp nov and description of *Carnobacterium inhibens* subsp *gilichinskyi* subsp nov, a psychrotolerant bacterium isolated from Siberian permafrost. *Int. J. Syst. Evol. Microbiol.* 65, 556–561. doi: 10.1099/ijss.0.067983-0
- Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Meth.* 8, 785–786. doi: 10.1038/nmeth.1701
- Pikuta, E. V. (2014). *The Family Carnobacteriaceae*, eds W. H. Holzapfel and B. J. B. Wood (Oxford: Blackwell Science Publishers). doi: 10.1002/9781118655252. part2
- Pikuta, E. V., and Hoover, R. B. (2014). *The Genus Carnobacterium*, eds W. H. Holzapfel and B. J. B. Wood (Oxford: Blackwell Science Publishers). doi: 10.1002/9781118655252.ch10
- Pilet, M. F., Dousset, X., Barré, R., Novel, G., Desmazaud, M., and Piard, J. C. (1994). Evidence for two bacteriocins produced by *Carnobacterium piscicola* and *Carnobacterium divergens* isolated from fish and active against *Listeria monocytogenes*. *J. Food Prot.* 58, 256–262. doi: 10.4315/0362-028X-58.3.256
- Proft, T., and Baker, E. N. (2009). Pili in Gram-negative and Gram-positive bacteria — structure, assembly and their role in disease. *Cell. Mol. Life Sci.* 66, 613–635. doi: 10.1007/s00018-008-8477-4
- Rahman, A., Cailliez-Grimal, C., Bontemps, C., Payot, S., Chaillou, S., Revol-Junelles, A.-M., et al. (2014a). High genetic diversity among strains of the unindustrialized lactic acid bacterium *Carnobacterium maltaromaticum* in dairy products as revealed by multilocus sequence typing. *Appl. Environ. Microbiol.* 80, 3920–3929. doi: 10.1128/AEM.00681-14
- Rahman, A., Gleinser, M., Lanham, M.-C., Riedel, C. U., Foligne, B., Hanse, M., et al. (2014b). Adaptation of the lactic acid bacterium *Carnobacterium maltaromaticum* LMA 28 to the mammalian gastrointestinal tract: from survival in mice to interaction with human cells. *Int. Dairy J.* 34, 93–99. doi: 10.1016/j.idairyj.2013.07.003
- Remenant, B., Borges, F., Cailliez-Grimal, C., Revol-Junelles, A.-M., Marché, L., Lajus, A., et al. (2016). Draft genome sequence of *Carnobacterium divergens* V41, a bacteriocin-producing strain. *Genome Announc.* 4:e01109-16. doi: 10.1128/genomeA.01109-16
- Rezaïki, L., Lamberet, G., Derré, A., Gruss, A., and Gaudu, P. (2008). *Lactococcus lactis* produces short-chain quinones that cross-feed Group B *Streptococcus* to activate respiration growth. *Mol. Microbiol.* 67, 947–957. doi: 10.1111/j.1365-2958.2007.06083.x
- Rodriguez, C., Taminiau, B., Brévers, B., Avesani, V., Van Broeck, J., Leroux, A., et al. (2015). Faecal microbiota characterisation of horses using 16 rRNA barcoded pyrosequencing, and carriage rate of *Clostridium difficile* at hospital admission. *BMC Microbiol.* 15:181. doi: 10.1186/s12866-015-0514-5
- Santagati, M., Campanile, F., and Stefani, S. (2012). Genomic diversification of enterococci in hosts: the role of the mobilome. *Front. Microbiol.* 3:95. doi: 10.3389/fmicb.2012.00095
- Savijoki, K., Ingmer, H., and Varmanen, P. (2006). Proteolytic systems of lactic acid bacteria. *Appl. Microbiol. Biotechnol.* 71, 394–406. doi: 10.1007/s00253-006-0427-1
- Schloss, P. D., and Handelsman, J. (2003). Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.* 14, 303–310. doi: 10.1016/S0958-1669(03)00067-3
- Schneewind, O., and Missiakas, D. (2014). Sec-secretion and sortase-mediated anchoring of proteins in Gram-positive bacteria. *Biochim. Biophys. Acta* 1843, 1687–1697. doi: 10.1016/j.bbamcr.2013.11.009
- Seedorf, H., Griffin, N. W., Ridaura, V. K., Reyes, A., Cheng, J., Rey, F. E., et al. (2014). Bacteria from diverse habitats colonize and compete in the mouse gut. *Cell* 159, 253–266. doi: 10.1016/j.cell.2014.09.008
- Siezen, R., Boekhorst, J., Muscariello, L., Molenaar, D., Renckens, B., and Kleerebezem, M. (2006). *Lactobacillus plantarum* gene clusters encoding putative cell-surface protein complexes for carbohydrate utilization are conserved in specific gram-positive bacteria. *BMC Genomics* 7:126. doi: 10.1186/1471-2164-7-126
- Siezen, R. J. (1999). Multi-domain, cell-envelope proteinases of lactic acid bacteria. *Antonie Van Leeuwenhoek* 76, 139–155. doi: 10.1023/A:1002036906922

- Sun, C., Fukui, H., Hara, K., Kitayama, Y., Eda, H., Yang, M., et al. (2015). Expression of Reg family genes in the gastrointestinal tract of mice treated with indomethacin. *Am. J. Physiol. Gastrointest. Liver Physiol.* 308, G736–G744. doi: 10.1152/ajpgi.00362.2014
- Tailford, L. E., Crost, E. H., Kavanaugh, D., and Juge, N. (2015). Mucin glycan foraging in the human gut microbiome. *Front. Genet.* 6:81. doi: 10.3389/fgene.2015.00081
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197
- Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., et al. (2013). MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.* 41, D636–D647. doi: 10.1093/nar/gks1194
- Voget, S., Klippe, B., Daniel, R., and Antranikian, G. (2011). Complete genome sequence of *Carnobacterium* sp 17-4. *J. Bacteriol.* 193, 3403–3404. doi: 10.1128/JB.05113-11
- Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Res.* 39, W347–W352. doi: 10.1093/nar/gkr485

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Iskandar, Borges, Taminiau, Daube, Zagorec, Remenant, Leisner, Hansen, Sørensen, Mangavel, Cailliez-Grimal and Revol-Junelles. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genetic Characterization of the Exceptionally High Heat Resistance of the Non-toxic Surrogate *Clostridium sporogenes* PA 3679

Robert R. Butler III¹, Kristin M. Schill², Yun Wang² and Jean-François Pombert^{1*}

¹ Department of Biology, Illinois Institute of Technology, Chicago, IL, USA, ² United States Food and Drug Administration, Center for Food Safety and Applied Nutrition, Bedford Park, IL, USA

Clostridium sporogenes PA 3679 is a non-toxic endospore former that is widely used as a surrogate for *Clostridium botulinum* by the food processing industry to validate thermal processing strategies. PA 3679 produces spores of exceptionally high heat resistance without botulinum neurotoxins, permitting the use of PA 3679 in inoculated pack studies while ensuring the safety of food processing facilities. To identify genes associated with this heat resistance, the genomes of *C. sporogenes* PA 3679 isolates were compared to several other *C. sporogenes* strains. The most significant difference was the acquisition of a second spoVA operon, spoVA2, which is responsible for transport of dipicolinic acid into the spore core during sporulation. Interestingly, spoVA2 was also found in some *C. botulinum* species which phylogenetically cluster with PA 3679. Most other *C. sporogenes* strains examined both lack the spoVA2 locus and are phylogenetically distant within the group I *Clostridium*, adding to the understanding that *C. sporogenes* are dispersed *C. botulinum* strains which lack toxin genes. *C. sporogenes* strains are thus a very eclectic group, and few strains possess the characteristic heat resistance of PA 3679.

OPEN ACCESS

Edited by:

Jennifer Ronholm,
McGill University, Canada

Reviewed by:

Bradley Stevenson,
University of Oklahoma, USA
Atte Von Wright,
University of Eastern Finland, Finland

*Correspondence:

Jean-François Pombert
jpombert@iit.edu

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 29 January 2017

Accepted: 15 March 2017

Published: 03 April 2017

Citation:

Butler RR III, Schill KM, Wang Y and Pombert J-F (2017) Genetic Characterization of the Exceptionally High Heat Resistance of the Non-toxic Surrogate *Clostridium sporogenes* PA 3679. *Front. Microbiol.* 8:545. doi: 10.3389/fmicb.2017.00545

Keywords: *Clostridium sporogenes*, *Clostridium botulinum*, PA 3679, SpoVA, dipicolinic acid, spore heat resistance, food sterilization, horizontal gene transfer

INTRODUCTION

Clostridium botulinum, *Clostridium baratii*, and *Clostridium butyricum* species produce various types of botulinum neurotoxin (BoNT), the causative agent of the neuroparalytic botulism poisoning (Dodds and Hauschild, 1989; Collins and East, 1998; Rossetto et al., 2014). These species cluster into six groups defined by their metabolic and physiological traits (Collins and East, 1998; Rossetto et al., 2014). Group I (proteolytic) *C. botulinum* strains are particularly important to the food industry, as they produce endospores of high heat resistance that may survive inadequate thermal processing strategies and result in food spoilage and foodborne botulism (Townsend et al., 1938; Gross et al., 1946; Ingram and Robinson, 1951; Stumbo et al., 1975; Rossetto et al., 2014). *Clostridium sporogenes* is closely related to *C. botulinum* group I strains, but differs in two characteristic respects: it lacks the BoNT toxin genes and it produces spores with even higher heat resistance (Nakamura et al., 1977; Bull et al., 2009; Brown et al., 2012; Diao et al., 2014).

C. sporogenes PA 3679 (PA 3679) is widely used in testing commercial thermal food processing procedures for their ability to prevent foodborne botulism in shelf-stable products (McClung, 1937; Brown et al., 2012; Rossetto et al., 2014). PA 3679 is a non-toxic surrogate possessing higher heat resistant spores than group I *C. botulinum*, providing a safe alternative test organism that ensures neurotoxic spores have been eliminated during the thermal process without introducing the target pathogen to the food processing facilities (Brown et al., 2012; Diao et al., 2014). PA 3679 was originally isolated from spoiled canned corn in 1927 by E.J. Cameron of the National Canner's Association (Townsend et al., 1938; Brown et al., 2012). However, the properties of PA 3679 that give it such high heat resistance have not been well explored at a genetic level.

Genes associated with spore heat resistance in endospore formers focus on three properties: (1) DNA damage prevention (and repair) (2) dipicolinic acid (DPA) and cation concentrations in the spore core and (3) the spore's core water content (Setlow, 2006, 2007, 2014b). Genes of particular interest are those under control of the sporulation sigma factors σ^G and σ^F (forespore-specific) or σ^E and σ^K (mother cell-specific) (Eichenberger et al., 2003; Molle et al., 2003; Huang et al., 2004; Dürre, 2005; Wang et al., 2006).

High temperatures destabilize DNA and increase the incidence of depurination (Setlow, 2007, 2014b). In metabolically suspended cells, accumulated mutations cannot be repaired until germination resumes cellular activity (Setlow, 2006, 2007, 2014b). In virtually all reported endospore formers, the presence of α/β -type small acid-soluble spore proteins (SASP) prevent this damage by binding to and stabilizing DNA in its A-form orientation (Setlow, 2007; Lee et al., 2008). This binding mechanism is suggested to require two conserved domains: a germination protease (*gpr*) cleavage domain, and a DNA-binding domain that facilitates the DNA-SASP adduct (Cabrera-Martinez and Setlow, 1991; Setlow, 2007; Lee et al., 2008; Wetzel and Fischer, 2015). This α/β -type SASP DNA protection method is highly conserved and so effective that wild-type spores that are killed by wet heat exhibit minimally damaged DNA, suggesting the disruption of some other spore component (Setlow, 2007, 2014b).

The concentration of DPA in the spore core, chelated in a 1:1 ratio with divalent cations (often Ca^{2+}), contributes to several spore functions (Granger et al., 2011; Setlow, 2014b). The magnitude of these effects can differ depending on the type of cation involved (Bach and Gilvarg, 1966; Paidhungat et al., 2000; Ragkousi et al., 2003; Setlow, 2014a,b). During sporulation, Ca^{2+} -DPA creates a high acidity, low water environment in the spore core and binds remaining free water therein (Paidhungat et al., 2000; Setlow, 2006, 2014b; Paredes-Sabja et al., 2008b; Donnelly et al., 2016). This core dehydration aids in DNA-SASP association, and more importantly prevents damage to spore proteins essential for revival and germination (Setlow, 2006, 2014a,b; Paredes-Sabja et al., 2008b).

DPA is synthesized in the mother cell by shunting the product of DapA, dihydrodipicolinic acid (DHDPA), from the process of lysine biosynthesis (Daniel and Errington, 1993; Orsburn et al., 2010). The dicistronic *spoVF* operon codes for

dipicolinic acid synthase subunits A and B (*dpaA/spoVFA* and *dpaB/spoVFB*), which convert DHDPA to DPA. Although the *spoVF* operon has been identified in many *Bacillus* species (Daniel and Errington, 1993; Onyenwoke et al., 2004) and *Peptoclostridium difficile* (Donnelly et al., 2016), this operon is not found in all endospore formers. Other members of the class Clostridia, including *Clostridium perfringens* and *Thermoanaerobacter* spp., lack *spoVF* (Onyenwoke et al., 2004). In *C. perfringens*, an alternate dipicolinic acid synthase, EtfA, has been demonstrated to produce DPA *in vitro* and *in vivo*, with knockout mutants lacking this metabolite (Orsburn et al., 2010). Following synthesis, DPA is transported to the core by three to seven products coded by the *spoVA* operon (Tovar-Rojo et al., 2002; Paredes-Sabja et al., 2008b; Li et al., 2012; Perez-Valdespino et al., 2014). Of these, products coded by *spoVAC*, *spoVAD* and *spoVAE* seem particularly important and are especially well conserved in both *Bacillus* and *Clostridium* species (Onyenwoke et al., 2004; Paredes-Sabja et al., 2008b; Donnelly et al., 2016).

In addition to DPA, several other genes are associated with core dehydration, though their roles are less clear. Spore maturation proteins A and B (products of *spmA* and *spmB*) both play a significant role in reducing core water content, though the mechanism is not understood (Paredes-Sabja et al., 2008a; Orsburn et al., 2009). The *dac* genes (*dacA*, *dacB*, *dacC*, and *dacF* in *B. subtilis*) code for D-alanyl-D-alanine carboxypeptidases which regulate peptidoglycan crosslinking. Both *dacB* and *dacF* genes are under the control of sporulation specific sigma factors, and their products regulate spore cortex formation (Popham et al., 1999). Knockout mutants lacking either gene show diminished heat resistance, presumably due to reduced cortex integrity under high heat conditions (Popham et al., 1999; Paredes-Sabja et al., 2008a; Orsburn et al., 2009).

In a previous study, we sequenced eight *C. sporogenes* samples labeled "PA 3679" obtained from a variety of sources and which displayed differential heat resistance (Schill et al., 2016). From our analyses, we distinguished two distinct clades of *C. sporogenes* isolates. Clade I isolates had significantly lowered heat resistance, with two (1990 and 2007) featuring near-identical genotypes. Clade I isolates did not survive heat treatment at 105°C for 5 min and displayed D_{97°C} and D_{100°C} values of 2.97 and 2.28 min, respectively (the decimal reduction time, D, is equal to the time required under a given condition to destroy a population of microorganisms by one logarithm). In contrast, all isolates from clade II exhibited near-identical genotypes and heat resistance profiles of the original PA 3679 isolate by E.J. Cameron, with an estimated D_{121°C} of 1.28 min (Diao et al., 2014), and survived thermal processing at temperatures from 117°C to 121°C. Given the two clades of *C. sporogenes* with differing heat resistance, we were presented with an opportunity to elucidate the specific genomic differences conferring the exceptional heat tolerance of PA 3679 spores.

MATERIALS AND METHODS

Genomes Used in Study

Eight genomes used in this study (Table 1) were from our previous study (Schill et al., 2016). The annotations of C.

TABLE 1 | *Clostridium sporogenes* isolates used in Schill et al. (2016).

Short name	Full name	Spore heat resistance group	Source
1961-2	<i>Clostridium sporogenes</i> 1961-2	clade I (low heat)	Contaminant of ATCC 7955 NCA3679
1990	<i>Clostridium sporogenes</i> 1990	clade I (low heat)	Contaminant of ATCC 7955 NCA3679
2007	<i>Clostridium sporogenes</i> 2007	clade I (low heat)	Contaminant of ATCC 7955 NCA3679
1961-4	<i>Clostridium sporogenes</i> PA 3679 1961-4	clade II (high heat)	ATCC 7955 NCA3679
Camp	<i>Clostridium sporogenes</i> PA 3679 Camp	clade II (high heat)	Campbell's Soup Company
FDA	<i>Clostridium sporogenes</i> PA 3679 FDA	clade II (high heat)	U.S. Food and Drug Administration
NFL	<i>Clostridium sporogenes</i> PA 3679 NFL	clade II (high heat)	National Food Laboratory
UW	<i>Clostridium sporogenes</i> PA 3679 UW	clade II (high heat)	Johnson Lab, University of Wisconsin-Madison

sporogenes 1961-2 (LLZW02), *C. sporogenes* 2007 (LLES02), *C. sporogenes* 1990 (LLZV01), *C. sporogenes* PA 3679 1961-4 (LLZT01), *C. sporogenes* PA 3679 Camp (LKKY02), *C. sporogenes* PA 3679 FDA (LJTA01), and *C. sporogenes* PA 3679 NFL (LJSZ01) and *C. sporogenes* PA 3679 UW (LFVV01) were updated to reflect the information from this study.

Pan-Genomic Analysis

For pan-genomic comparison, the eight strains previously described were clustered using Roary 3.6.2 (Page et al., 2015) using a 70% identity threshold. Roary's core gene alignment was trimmed using BMGE 1-1 (Criscuolo and Gribaldo, 2010) to 2,511,737 sites across 2,751 core genes. PhyML 3.1 (Guindon et al., 2010) was used with GTR + I + F + G (4 categories) to generate a maximum likelihood (ML) tree for clustering. Roary_plots.py (https://github.com/sanger-pathogens/Roary/tree/master/contrib/roary_plots) was used to generate the orthologous cluster map. Orthologs unique to the five clade II isolates were examined and those related to sporulation were investigated.

Roary was used with 23 *C. sporogenes* (including the eight in this study) and 15 group I *C. botulinum* genomes, to generate a concatenated nucleotide alignment of 389 core genes (216,294 sites) using a 70% identity threshold. A maximum likelihood tree was generated as above, with the addition of 100 bootstraps in PhyML.

To calculate pairwise mutational distances between the 38 group I *Clostridium*, Mash (Ondov et al., 2016) pairwise comparisons were plotted using metric multidimensional scaling with the cmdscale and igraph (Csárdi and Nepusz, 2006) packages implemented in R (R Core Team, 2016), using custom Perl scripts available via the Pombert Lab github page (<https://github.com/PombertLab>).

Analysis of *spoVA* and Conserved Genes

Orthologous groups identified from the pan-genomic analysis were searched using known genes related to spore heat resistance. Additional homology searches using BLAST (Altschul et al., 1990) looked for any missed homologs. Both blastp and tblastn searches were conducted using known reference genes (See Supplementary Table 1). Orthologous groups for each gene were compared and aligned using Geneious 9.1.5 (Kearse et al., 2012). Conserved domains in the aligned clusters were revealed with InterProScan 5 (Jones et al., 2014).

Analysis of the *spoVA* Operon

Using the 38 *Clostridium* above, plus *C. tetani* E88, OrthoFinder 0.2.8 (Emms and Kelly, 2015) identified 6,168 orthologous groups, 840 of which were unique orthologs present in all 39 strains. All *spoVA* genes identified by the pan-genomic analysis were located in the ortholog groups produced by OrthoFinder. Neighboring genes and operons were also identified in all 38 group I *Clostridium* species examined. *spoVA2* operons and neighboring genes for several representative species were aligned and compared using EasyFig 2.2.2 (Sullivan et al., 2011). Conserved domains were identified using InterProScan 5, and predicted protein structures were calculated using the RaptorX webserver (Källberg et al., 2012).

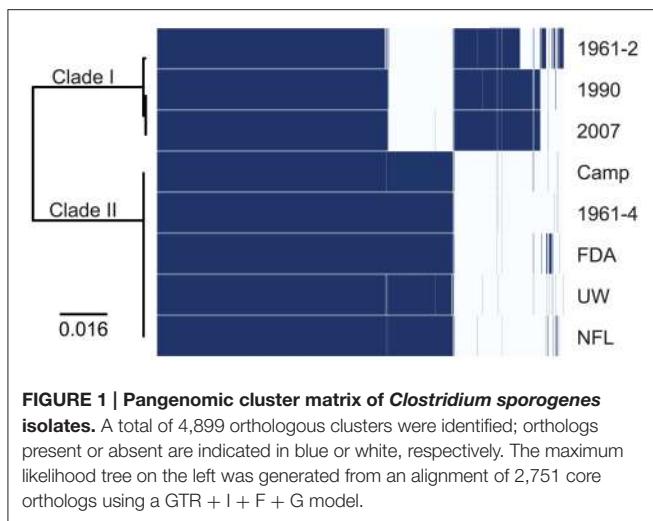
RESULTS

Pan-Genome Analysis

The Roary pan genome generated using the eight clade I and clade II *C. sporogenes* isolates contained a total of 4,899 distinct orthologous groups, 2,751 of which represented the core genes, each with a unique ortholog in all eight isolates (Figure 1). These core orthologs included many genes previously identified as related to spore heat resistance. The heat resistance core orthologs are further characterized in Supplementary Figure 1, and described later in detail. There were 751 ortholog groups that were present in the five clade II isolates, but absent in clade I. Of those, 278 ortholog groups code for hypothetical proteins. Seven of the 751 were sporulation specific, of which four ortholog groups were related to germination. The remaining three ortholog groups constituted a second set of *spoVA* genes not found in the clade I isolates, henceforth dubbed the *spoVA2* locus.

Examination of the *spoVA2* operon

As mentioned in the pan genomic analysis, a second locus of *spoVA* genes was found in a single pentacistronic operon, *spoVA2* (Figure 2A). InterProScan searches of the *spoVA2* genes revealed conserved domains from two types of *spoVA* operons. To explore this further, a collection of the *C. sporogenes* genomes in GenBank (at the time of writing) plus fifteen commonly studied group I *C. botulinum* strains and one *C. tetani* strain (Table 2) were clustered using OrthoFinder. For all five clade II isolates and five additional *C. botulinum* species, this *spoVA2* operon was conserved and clustered separately



from the traditional *spoVA* operon, which was found in all 39 species. All 39 species showed similar *spoVA* loci as clustered in Orthofinder. The 38 group I Clostridia showed a conserved genomic neighborhood around the site of the *spoVA2* operon inclusion (**Figure 2B**). The *spoVA2* operon and its neighboring regions were perfectly conserved in all clade II (PA 3679) isolates, so only one representative sequence (Camp) is depicted in the figure. The clade I isolates similarly only had a single nucleotide difference across the whole region which didn't affect gene coding, so 2007 was chosen as the representative in **Figure 2B**.

The *spoVA2* operon itself was well conserved in all strains it was found in. In addition to SpoVAC, SpoVAD, and SpoVAE_b, the operon encodes two other proteins: a hypothetical protein and a membrane protein (**Figure 2A**). Neither protein has domain similarity or sequence homology to SpoVAA or SpoVAB. Both feature a domain of unknown function (DUF), DUF1657 (IPR012452; PF07870), and the membrane protein contains an additional uncharacterized YcaP domain (PTHR34582) composed of three transmembrane domains and DUF421. Predicted 3D structures of the DUF1657 hypothetical protein, the YcaP/DUF1657 membrane protein and SpoVAC, SpoVAD, and SpoVAE are depicted in Supplementary Figure 2, with high similarity to previously reported examples. Also of note is the downstream neighbor of *spoVA2*, the xanthine dehydrogenase (*xdh*) operon. The *xdh* operon and a gene encoding isochorismate hydrolase are present in all *spoVA2* containing strains, as well as closely related *C. botulinum* A strain ATCC 3502 (which lacks *spoVA2*; **Figure 2B**). However, in several strains, the *xdh* genes are partial or pseudogenes.

Characterization of SASP

A total of eight different SASP-encoding ortholog groups were found, each group containing an ortholog from every one of the eight investigated genomes. The traditional α/β -type SASP, with both a *gpr* cleavage domain (IPR018126; Prosite PS00304) and a DNA-binding domain

(IPR018126; Prosite PS00684), was encoded by three of these orthologous groups. Translations of the genes in those groups, named *ssp1*, *ssp2*, and *ssp3*, also displayed the characteristic α/β -type SASP Pfam domain (IPR001448; PF00269).

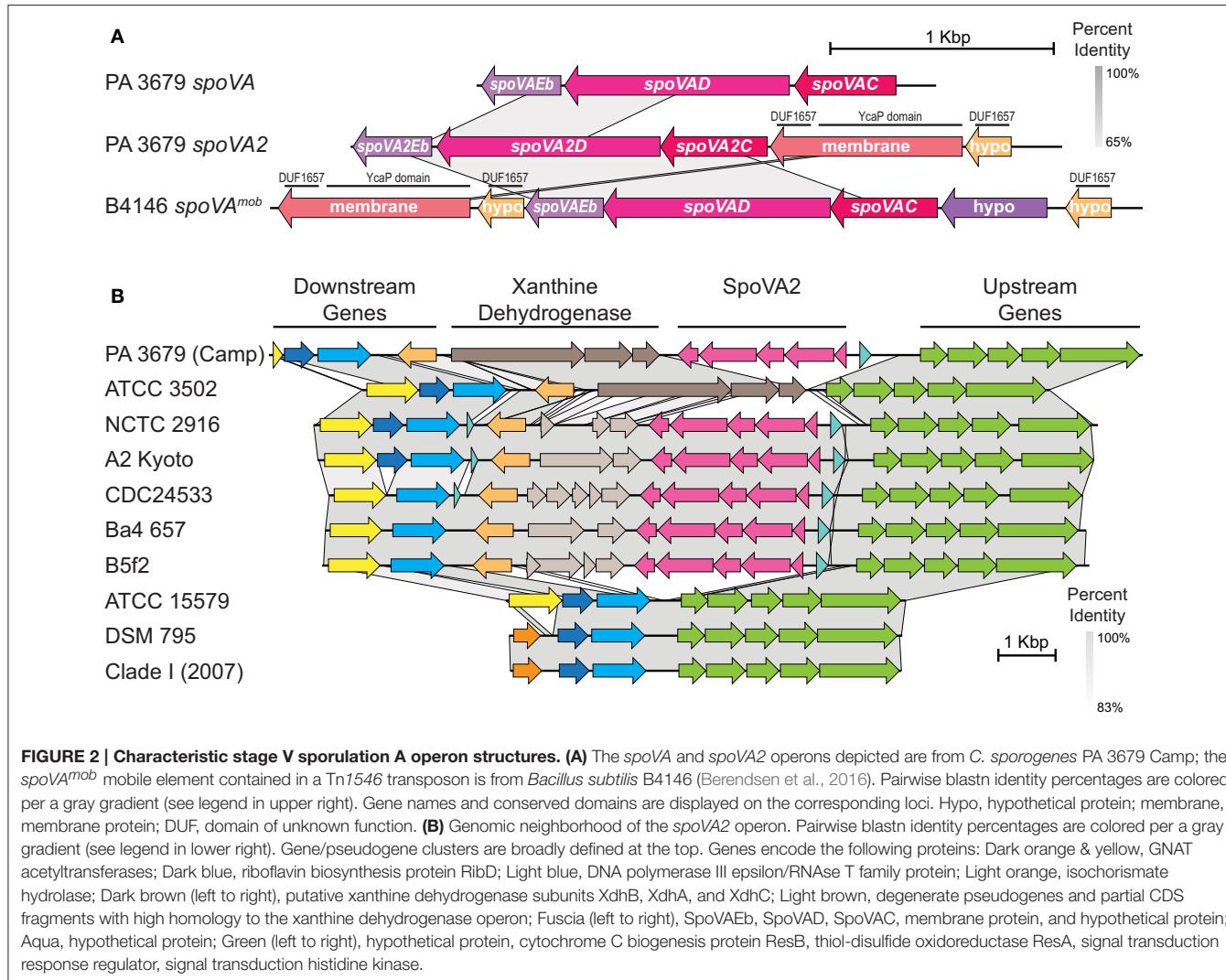
A fourth SASP-encoding ortholog group showed high sequence conservation to a previously described *ssp4* in *C. perfringens* (Li and McClane, 2008; Li et al., 2009). The product coded by these *ssp4* orthologs had the characteristic α/β -type SASP Pfam domain, and the *gpr* cleavage domain, but lacked the conserved DNA-binding domain. The fifth SASP-encoding ortholog group contained orthologs labeled *ssp5*. Again, translations displayed the conserved SASP Pfam domain, however it lacked both the *gpr* cleavage domain and the DNA-binding domain typical of α/β -type SASP.

The remaining three SASP-encoding ortholog groups exhibited the conserved domains and sequence similarity to minor types of SASP not associated with high heat resistance in previous studies: the H-type SASP and the *tpl* type SASP (Cabrera-Hernandez et al., 1999; Wetzel and Fischer, 2015). All of the SASP-encoding ortholog groups in this study appear to be monocistronic, and the amino acid alignments of each SASP is available in Supplementary Figure 1.

Characterization of Conserved Sporulation Genes

A number of additional sporulation-related orthologous groups were found with representative orthologs from all eight isolates. Six D-alanyl-D-alanine carboxypeptidase encoding orthologous groups were found, and their respective orthologous genes were dubbed *dac1* through *dac6*. One orthologous group, *dac4*, encoded proteins with high homology to DacF (blastp e-values above 1e-105) and contained the two expected conserved domains: Peptidase S11, N-terminal domain (IPR001967; Pfam PF00768) and Penicillin Binding Protein 5, C-terminal domain (PBP5_C) (IPR012907; Pfam PF07943). Two orthologous groups—*dac2* and *dac5*—encoded proteins similar to DacB (with blastp e-values above 1e-49) which characteristically have the same two conserved domains as DacF. The three remaining *dac* ortholog groups (*dac1*, *dac3*, and *dac6*) showed poor similarity to *dacB* or *dacF* and are likely D-alanyl-D-alanine carboxypeptidases unrelated to sporulation. The amino acid alignments of all Dac proteins are available in Supplementary Figure 1.

The *spoVF* operon, coding for DPA synthase subunits A and B, was not found in any of the eight isolates. Instead, three orthologous groups—containing orthologs dubbed *etfA_1*, *etfA_2*, and *etfA_3*—encoded products with high protein sequence similarity (blastp e-values above 1e-130) to EtFA from *C. perfringens*, an alternate DPA synthase. All three EtFA homologs were present in all eight genomes. Only EtFA_1 contained the correct array of conserved domains associated with *C. perfringens* EtFA. EtFA_3 lacked a Prosite conserved motif (IPR018206; PS00696) and EtFA_2 contained an extra conserved domain: N-terminal 4Fe-4S ferredoxin-type iron-sulfur binding



domain (FerB) (IPR017896; Pfam PF00037). The amino acid alignments of the EtfA proteins are available in Supplementary Figure 1.

The orthologous groups for 4-hydroxy-tetrahydrodipicolinate synthase (DapA) and 4-hydroxy-tetrahydrodipicolinate reductase (DapB) were both present in all eight isolates, as was a second DapB orthologous group (encoded by *dapB_2*). Orthologous groups encoding spore maturation protein A (SpmA) and B (SpmB) were also found and contained an ortholog in all eight isolates. Other orthologs typically associated with germination were also identified in the isolates. Germination protease (*gpr*), putative germination protease (*yyaC*) and spore photoproduct lyase (*splB*) orthologs were also found in all eight genomes. Supplementary Table 1 summarizes the orthologous genes and includes those which were found in *C. botulinum* A strain ATCC 3502. Locus tags and further information for all the genes in this study can be found in Supplementary Table 1.

Phylogenomic Comparison

The phylogeny in Figure 3 (upper) depicts a branching of *C. sporogenes* and *C. botulinum* strains into two mixed groups. The majority of *C. sporogenes* strains are grouped together in the right group, though clade II (PA 3679) isolates group on the left. The majority of *C. botulinum* strains are in the left group, though several are present in the right group. All strains possessing the *spoVA2* locus are in the left group. The xanthine dehydrogenase operon and isochorismate hydrolase are present in all members of the left group, though degenerated in some strains, and absent in all strains in the right group. The clade II isolates form an extremely well conserved group consistent with coming from the same original spore crop. The clade I isolates also group as expected, showing similarity to several *C. sporogenes* strains and one *C. botulinum* strain, Prevot 594. The pairwise genetic distances comparison in Figure 3 (lower) shows consistent results with the core gene phylogenetic tree, however the species in the right branch of the phylogeny are split into

TABLE 2 | Group I Clostridia used in spoVA2 comparisons.

Strain name	Species	Toxin type	Accession
AM1195	<i>Clostridium botulinum</i>	B6	LFPH01
ATCC 19397	<i>Clostridium botulinum</i>	A1	CP000726.1
ATCC 3502	<i>Clostridium botulinum</i>	A1	AM412317.1
B5f2	<i>Clostridium botulinum</i>	B5f2	ABDP01
Ba4 657	<i>Clostridium botulinum</i>	B5a4	CP001083.1
F230613	<i>Clostridium botulinum</i>	F1	CP002011.1
H04402-065	<i>Clostridium botulinum</i>	A5	FR773526.1
Hall	<i>Clostridium botulinum</i>	A1	CP000727.1
Kyoto	<i>Clostridium botulinum</i>	A2	CP001581.1
Langeland	<i>Clostridium botulinum</i>	F1	CP000728.1
Loch Maree	<i>Clostridium botulinum</i>	A3	CP000962.1
NCTC 2916	<i>Clostridium botulinum</i>	A1(B)	ABDO02
Okra	<i>Clostridium botulinum</i>	B1	CP000939.1
Osaka05	<i>Clostridium botulinum</i>	B6	BAUF01
Prevot 594	<i>Clostridium botulinum</i>	B	CP006902.1
11579	<i>Clostridium sporogenes</i>	–	JZJN01
66_CBOT	<i>Clostridium sporogenes</i>	–	JUYE01
85-3852	<i>Clostridium sporogenes</i>	–	JZJO01
87-0535	<i>Clostridium sporogenes</i>	–	JZJP01
88-0163	<i>Clostridium sporogenes</i>	–	JZJQ01
8-O	<i>Clostridium sporogenes</i>	–	LUAU01
ATCC 15579	<i>Clostridium sporogenes</i>	–	ABKW02
ATCC 19404	<i>Clostridium sporogenes</i>	–	LFPM01
Bradbury	<i>Clostridium sporogenes</i>	–	AGAH01
CDC23284	<i>Clostridium sporogenes</i>	–	LAGF01
CDC24533	<i>Clostridium sporogenes</i>	–	LAGH01
DSM 795 [†]	<i>Clostridium sporogenes</i>	–	CP011663.1
DSM 795	<i>Clostridium sporogenes</i>	–	JFBQ01
NCIMB 10696	<i>Clostridium sporogenes</i>	–	CP009225.1
UC9000	<i>Clostridium sporogenes</i>	–	LJFK01
E88 ^a	<i>Clostridium tetani</i>	–	AE015927.1

^aNot a group I *Clostridium* species, outgroup used for clustering and phylogeny. [†]Two distinct whole genome submissions from different groups for strain DSM 795 are available in GenBank. The cross identifies which strain corresponds to which accession number in **Figure 3**.

two more distinct groups than the phylogenetic tree alone would suggest.

DISCUSSION

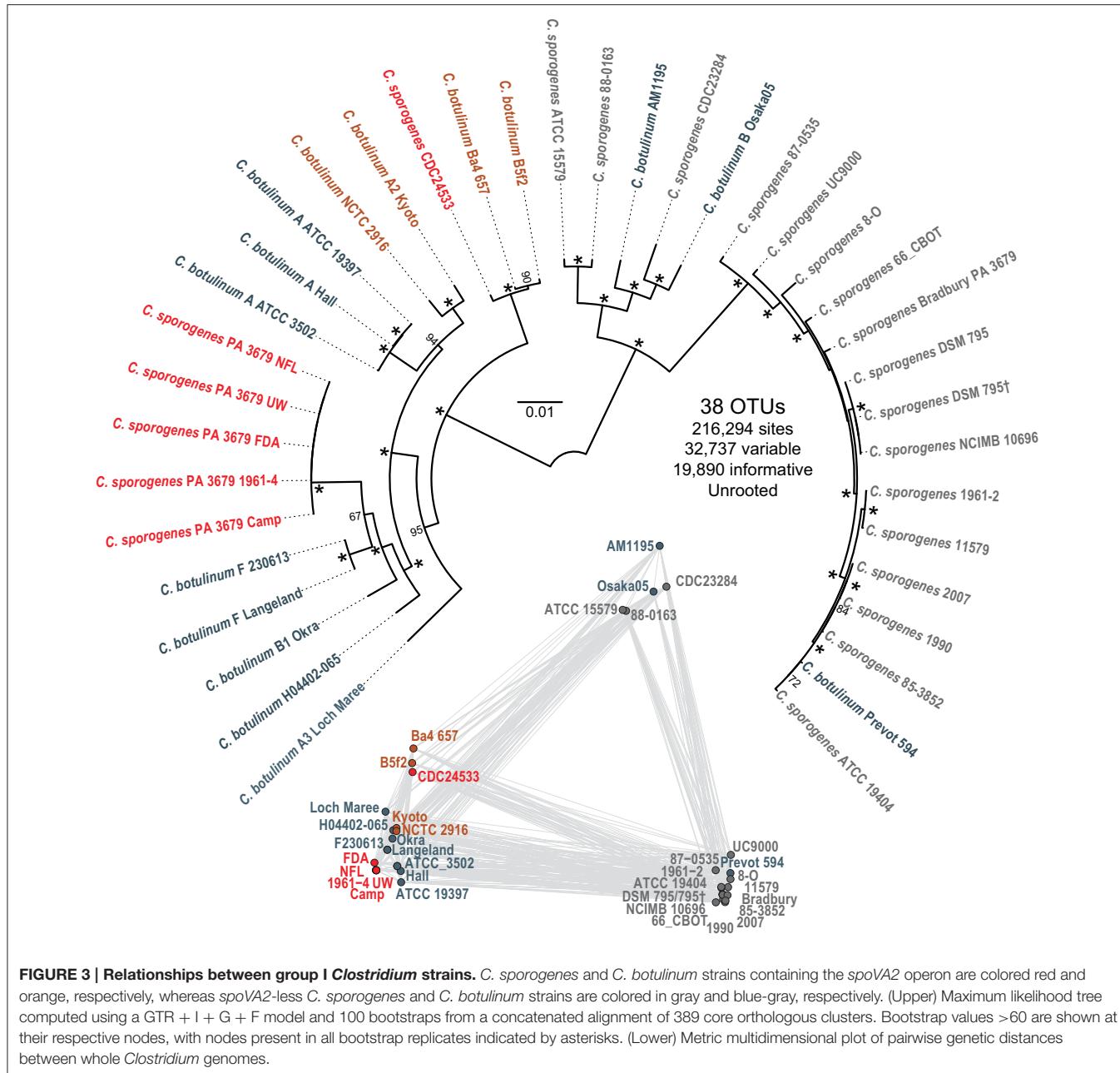
Ensuring the quality and safety of packaged foods is an ongoing process that is ideally unnoticed by the consumer when everything works as intended. While current methods for food preservation have an excellent track record, bacteria do evolve over time and there is a non-negligible risk that these methods may no longer be adequate in the near future (for example, the rise of antibiotic resistance is a sharp reminder that things can change quickly in the microbial world). Here, we were offered the opportunity to examine the genetic differences behind the low and high heat resistance of clade I and II isolates of *C. sporogenes*; a knowledge that could be applied to

the detection and prevention of heat resistance in pathogenic species of *Clostridium* and other common foodborne pathogens. Importantly, we discovered that the genetic locus that most likely conveys a meaningful improvement in PA 3679 spore heat resistance is part of the bacterial mobilome, and that this heat resistance-conferring island could be, and likely has been, transferred from/to a number of pathogenic species. Our study also serves as a reminder that not all *C. sporogenes* isolates identified as PA 3679 strains actually possess the capability for high heat resistance and therefore cannot fulfill the role of non-toxic, thermal surrogates. Processes vetted using these deficient strains may not perform up to the desired specifications, with potentially dire implications for the safety of the foods packaged by these processes.

Impact of the spoVA2 Locus

Undoubtedly, the most significant difference found between the low and high heat resistance isolates was the presence of a second set of *spoVA* genes in the clade II (PA 3679) group. This is not the first reported incidence of multiple *spoVA* operons in an endospore former. At least one group IV *Clostridium* species (Brunt et al., 2016), several species of *Geobacillus* and *Bacillus cereus*, as well as several of the more heat resistant *Bacilli* have multiple *spoVA* loci. Adding additional copies of *spoVA* in a mobile Tn1546 transposon (*spoVA*^{mob}, **Figure 2A**) creates an additive resistance effect, greatly increasing the concentration of DPA in the *Bacillus subtilis* spore core (a D_{112.5°C} increase from 0.2 min to 25.6 min with three copies; Berendsen et al., 2016). Our results are compatible with these previous observations, and lend to a compelling hypothesis about spore heat resistance. As spore formation is temporally limited, the SpoVA apparatus encounters a flow rate challenge. Increasing the flow rate with extra pumps can either move more DPA in the given time, or overcome losses due to diffusion (or both) resulting in a much higher concentration of DPA in the spore core. This effect should scale until the point that the maximal amount of DPA has been added, which has apparently not been reached in *Bacillus*, and our findings suggest the same for Clostridia, with the implication that further multiplication of this operon in the genetic paraphernalia of an endospore former may imbue it with the ability to survive current canning processes.

The origin of the PA 3679 *spoVA2* operon, however, is not entirely clear. This operon was not contained in the same Tn1546 mobile element as in *Bacillus* species, but individual genes within the *spoVA2* locus—hypothetical protein, membrane protein, *spoVAC*, *spoVAD*, and *spoVAEB*—showed a higher sequence similarity to foreign loci than to the native *spoVA* locus in PA 3679 (**Figure 2A**). Blastn searches of the contiguous *spoVA2* operon gave high sequence homology (>80% identity, >99% query coverage) to the expected *C. botulinum* species from this study, plus *C. argentinense* CDC 2741, *C. neonatale*, *C. saccharobutylicum* DSM 13864, and *C. saccharoperbutylacetoni* N1-4. Given this information, horizontal acquisition seems more likely than a paralogous duplication event. This idea is furthered when considering the two additional genes (coding for the DUF1657 domain-containing and YcaP domain-containing proteins) which show a homology to *spoVA2*^{mob} from *B. subtilis* yet are



absent in the native *spoVA* locus. The YcaP domain-containing protein is of particular interest as Berendsen et al. (2016) knocked out the orthologous protein in *spoVA*^{mob}, which severely diminished spore heat resistance in *B. subtilis*. While the roles of SpoVAD (Li et al., 2012) and SpoVAC (Velásquez et al., 2014) are partially established in DPA transport, a full understanding of the roles of all *spoVA2* proteins needs further study.

The presence of additional *spoVA* operons in *Clostridium* species has not been previously explored, though it is a phenomenon that occurs not just in PA 3679, but also in several closely related *C. botulinum* species (Figure 2B) and at least one *C. argentinense* strain (Brunt et al., 2016). This might seem paradoxical as *C. botulinum* is generally considered to have

lower spore heat resistance than *C. sporogenes*. However, (1) there is a large amount of variance and inconsistency in heat resistance data, owing to a variety of environmental factors involved, (2) the heat resistance of the *C. botulinum* species possessing the *spoVA2* locus has not been widely studied and (3) this study only examined a comparison of *C. sporogenes* strains. Perhaps *C. botulinum* species containing the *spoVA2* locus also feature increased heat resistance. This would present a considerable challenge designing thermal processing strategies which effectively eliminate this dangerous pathogen in a food product. Future studies will be required to explore other spore heat resistance factors that may differ between the *C. botulinum* and *C. sporogenes*.

High Conservation of SASPs

The α/β -type SASP have been recognized primarily for their function in maintaining spore DNA integrity when exposed to a variety of factors (Setlow, 2014b). Many studies of α/β -type SASP knockouts have demonstrated a significant loss of heat resistance when lacking a functional DNA protection mechanism (Setlow, 2007, 2014b). However, the protection provided by SASP is not additive, as only one or two of the paralogous SASP-encoding genes are expressed in large amounts, and confer maximal heat resistance (Setlow, 2014b). The presence of eight SASP-encoding orthologous groups in our isolates clouded the search for differential heat resistance, thus it was decided to focus on faulty or absent SASP. The eight isolates in our study share three orthologs similar to those in other Clostridia: *ssp1*, *ssp2*, and *ssp3* (Raju et al., 2006; Galperin et al., 2012). These encode proteins containing all the major α/β -type SASP conserved domains, and show very little difference in protein sequence between clade I and clade II isolates, suggesting the presence of functional SASP-DNA protection in all isolates (Supplementary Figure 1). Additionally, the minor SASP—*ssp5*, H-type, and *tlp*—are also well conserved though not directly implicated in heat resistance.

The one SASP that demonstrated a unique feature was that encoded by the *ssp4* orthologous group. Previous research on the orthologous SASP in *C. perfringens* suggested that the presence of an aspartate (D), or other negatively charged or large amino acid at position 36 (Li and McClane, 2008) correlates with higher heat resistance when compared to other residues (Li et al., 2009). The clade I and II strains in this study displayed either a threonine (T) or isoleucine (I) residue at this position, respectively; and it is worth noting that *C. botulinum* A strain ATCC 3502 features an Ssp4 with an I at that position, yet still produces spores with a lower heat resistance than PA 3679. The lack of a negative charge at this position also does not appear to impede spore heat resistance for clade II (PA 3679) isolates. While a potential increase in spore heat resistance for PA 3679 with an I36D mutation is worth investigating, it would appear that the SASP-DNA protection mechanism provided by Ssp1-3 is already sufficient given its current robustness. As Setlow (2014b) has suggested, this dynamic hits a saturation point, beyond which more or better SASPs are no longer the limiting factor for higher spore heat resistance and the potential effect, if any, of T36I or T36D substitutions in Ssp4 for improving heat resistance of clade I isolates is unclear.

Conserved Sporulation Genes

Possessing six D-alanyl-D-alanine carboxypeptidases appears typical for many *Clostridium* and *Bacillus* species. All eight isolates in this study have potential orthologs for DacB and DacF, the two carboxypeptidases which have a demonstrated effect on spore heat resistance. Dac2 through Dac5 contained all the expected conserved domains. Based on sequence homology, Dac4 is most likely the DacF ortholog, and the DacB homolog is likely either Dac2 or Dac5 (both showed similar *e*-values). Ultimately, the determination of the role of each Dac will require future experiments to determine which one is regulated by σ^F

(DacF, expressed in the forespore) and which one by σ^E (DacB, expressed in the mother cell). From this study, none of the potential orthologs appeared to be significantly different between the clade I and clade II isolates (Supplementary Figure 1), thus they are likely not responsible for the differential heat resistance.

DPA synthesis in these *C. sporogenes* isolates is not controlled via a *spoVF* mechanism, though potentially is synthesized via an electron transport flavoprotein α -subunit as seen in *C. perfringens* (Orsburn et al., 2010). The EtfA_3 orthologous group lacked a C-terminal Prosite conserved domain, and EtfA_2 orthologous group had an extraneous N-terminal FerB domain, making them both unlikely candidates. The EtfA_1 product, which contained all the expected domains and had the highest sequence homology, is the most likely ortholog. Future experiments will need to replicate the experiments from Orsburn et al. (2010) in order to prove conclusively that the product of *etfA_1* is capable of DPA synthesis *in vitro* and *in vivo*. An electron transport flavoprotein is common and this phenomenon is fairly unique. Regardless, the three potential orthologous groups show a high degree of sequence conservation in all eight isolates, thus none of them are likely to account for the heat resistance difference we see between clade I and clade II isolates.

The SpmA and SpmB orthologs were present and highly conserved, generating little ambiguity about their identities. All expected domains were present, and minimal variation between clade I and clade II sequences make it unlikely that they contribute to the differential heat resistance. All additional genes examined showed a very high sequence similarity to unique orthologous groups containing representatives from all eight isolates. Plus, their involvement in spore heat resistance is mostly tangential, again making them unlikely factors in the observed change (For more information see Supplementary Table 1 and Supplementary Figure 1).

Interrelatedness of Group I *Clostridium* Species and Origins of the *spoVA2* Locus

The phylogenetic tree produced in this study (Figure 3) was consistent with previous studies (Kenri et al., 2014; Weigand et al., 2015; Williamson et al., 2016). However, PA 3679 strains did not group with other *C. sporogenes* strains, instead clustering deep in the left branch. The Mash pairwise distances corroborated the phylogenetic tree, demonstrating not only the position of PA 3679 strains with a group of *C. botulinum*, but heterogeneity among *C. sporogenes* strains in general. Considering how very different PA 3679 strains appear from the other *C. sporogenes* strains, the assertion that *C. sporogenes* strains in general are suitable non-toxic surrogates is questionable. Most of the *C. sporogenes* strains examined lack the *spoVA2* locus, and given their phylogenetic relatedness to the clade I isolates from this study, it is likely that they possess similar low heat resistance profiles. *C. sporogenes* CDC24533 is the exception—possessing *spoVA2*—and has the potential to produce spores that are resistant to high temperatures similar to PA 3679, warranting further investigation.

The high degree of conservation observed between the *spoVA2* operons present in species from the left side of the tree (Figure 3) argues in favor of a common origin, but it is unclear if this distribution results from multiple independent acquisition events from similar sources or rather from a single acquisition in their shared common ancestor followed by independent losses. While heat resistance confers an obvious advantage to species exposed to extreme temperatures like those involved in canning processes, those conditions are rarely met in the environment and one can envision that the added benefit may be rather minimal in normal circumstances, and thus commonly lost during pruning processes. In any case, the presence of the *spoVA2* operon in botulinum neurotoxin-containing *Clostridium* species strongly argues in favor of maintaining stringent canning processes that meet or exceed spore destruction targets of heat-resistant *C. sporogenes* isolates.

CONCLUSIONS

The high heat resistance of *Clostridium sporogenes* PA 3679 is unique among observed *C. sporogenes* strains. While this resistance is most likely influenced by the presence of an extra *spoVA2* operon, other factors including differential expression, altered function of canonical sporulation proteins and/or additional novel sporulation proteins could be involved. Further, studies will be required to circumscribe the full set of factors that confer to PA 3679 this thermal endurance and to better define the mechanisms that are involved in its endospore survival. Furthermore, because the potential for higher heat resistance also exists in both harmless and pathogenic species, strategies to

detect and reduce thermal stability in foodborne organisms as well as to how maintain safe standards of food processing will need to be revisited.

AUTHOR CONTRIBUTIONS

RB, JP, KS, and YW designed the study and drafted the manuscript. RB conducted the work and RB and JP conducted the analysis.

FUNDING

This work was supported by a C.V. Starr fellowship to RB and by funds from the Illinois Institute of Technology to JP and YW was supported by an appointment to the Research Participation Program at the Center for Food Safety and Applied Nutrition administered by the Oak Ridge Institute for Science and Education via an interagency agreement between the U.S. Department of Energy and the FDA.

ACKNOWLEDGMENTS

We would like to thank Yukun Sun and Iva Veseli for assistance with Perl and R scripts.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.00545/full#supplementary-material>

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2
- Bach, M. L., and Gilvarg, C. (1966). Biosynthesis of dipicolinic acid in sporulating *Bacillus megaterium*. *J. Biol. Chem.* 241, 4563–4564.
- Berendsen, E. M., Boekhorst, J., Kuipers, O. P., and Wells-Bennik, M. H. (2016). A mobile genetic element profoundly increases heat resistance of bacterial spores. *ISME J.* 10, 2633–2642. doi: 10.1038/ismej.2016.59
- Brown, J. L., Tran-Dinh, N., and Chapman, B. (2012). *Clostridium sporogenes* PA 3679 and its uses in the derivation of thermal processing schedules for low-acid shelf-stable foods and as a research model for proteolytic *Clostridium botulinum*. *J. Food Prot.* 75, 779–792. doi: 10.4315/0362-028X.JFP-11-391
- Brunt, J., van Vliet, A. H., van den Bos, F., Carter, A. T., and Peck, M. W. (2016). Diversity of the germination apparatus in *Clostridium botulinum* groups I, II, III, and IV. *Front. Microbiol.* 7:1702. doi: 10.3389/fmicb.2016.01702
- Bull, M. K., Olivier, S. A., van Diepenbeek, R. J., Kormelink, F., and Chapman, B. (2009). Synergistic inactivation of spores of proteolytic *Clostridium botulinum* strains by high pressure and heat is strain and product dependent. *Appl. Environ. Microbiol.* 75, 434–445. doi: 10.1128/AEM.01426-08
- Cabrera-Hernandez, A., Sanchez-Salas, J.-L., Paidhungat, M., and Setlow, P. (1999). Regulation of four genes encoding small, acid-soluble spore proteins in *Bacillus subtilis*. *Gene* 232, 1–10. doi: 10.1016/S0378-1119(99)00124-9
- Cabrera-Martinez, R. M., and Setlow, P. (1991). Cloning and nucleotide sequence of three genes coding for small, acid-soluble proteins of *Clostridium perfringens* spores. *FEMS Microbiol. Lett.* 61, 127–131. doi: 10.1111/j.1574-6968.1991.tb04335.x
- Collins, M. D., and East, A. K. (1998). Phylogeny and taxonomy of the food-borne pathogen *Clostridium botulinum* and its neurotoxins. *J. Appl. Microbiol.* 84, 5–17. doi: 10.1046/j.1365-2672.1997.00313.x
- Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* 10:210. doi: 10.1186/1471-2148-10-210
- Csárdi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Inter J. Complex Syst.* 1695, 1–9. Available online at: http://interjournal.org/manuscript_abstract.php?361100992
- Daniel, R. A., and Errington, J. (1993). Cloning, DNA sequence, functional analysis and transcriptional regulation of the genes encoding dipicolinic acid synthetase required for sporulation in *Bacillus subtilis*. *J. Mol. Biol.* 232, 468–483. doi: 10.1006/jmbi.1993.1403
- Diao, M. M., André, S., and Membré, J.-M. (2014). Meta-analysis of D-values of proteolytic *Clostridium botulinum* and its surrogate strain *Clostridium sporogenes* PA 3679. *Int. J. Food Microbiol.* 174, 23–30. doi: 10.1016/j.ijfoodmicro.2013.12.029
- Dodds, K. L., and Hauschild, A. H. (1989). “Distribution of *Clostridium botulinum* in the environment and its significance in relation to botulism,” in *Proceedings of the 5th International Symposium on Microbial Ecology* (Kyoto: International Society for Microbial Ecology), 472.
- Donnelly, M. L., Fimlaid, K. A., and Shen, A. (2016). Characterization of *Clostridium difficile* spores lacking either *SpoVA* or dipicolinic acid synthetase. *J. Bacteriol.* 198, 1694–1707. doi: 10.1128/JB.00986-15
- Dürre, P. (2005). “Sporulation in clostridia (Genetics),” in *Handbook on Clostridia*, ed P. Dürre (Boca Raton, FL: CRC Press), 659–669.
- Eichenberger, P., Jensen, S. T., Conlon, E. M., van Ooij, C., Silvaggi, J., González-Pastor, J.-E., et al. (2003). The σ^E regulon and the identification of

- additional sporulation genes in *Bacillus subtilis*. *J. Mol. Biol.* 327, 945–972. doi: 10.1016/S0022-2836(03)00205-5
- Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157. doi: 10.1186/s13059-015-0721-2
- Galperin, M. Y., Mekhedov, S. L., Puigbo, P., Smirnov, S., Wolf, Y. I., and Rigden, D. J. (2012). Genomic determinants of sporulation in Bacilli and Clostridia: towards the minimal set of sporulation-specific genes. *Environ. Microbiol.* 14, 2870–2890. doi: 10.1111/j.1462-2920.2012.02841.x
- Granger, A. C., Gaidamakova, E. K., Matrosova, V. Y., Daly, M. J., and Setlow, P. (2011). Effects of Mn and Fe levels on *Bacillus subtilis* spore resistance and effects of Mn²⁺, other divalent cations, orthophosphate, and dipicolinic acid on protein resistance to ionizing radiation. *Appl. Environ. Microbiol.* 77, 32–40. doi: 10.1128/AEM.01965-10
- Gross, C. E., Vinton, C., and Stumbo, C. R. (1946). Bacteriological studies relating to thermal processing of canned meats. V. Characteristics of putrefactive anaerobe used in thermal resistance studies. *J. Food Sci.* 11, 405–410. doi: 10.1111/j.1365-2621.1946.tb16368.x
- Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010
- Huang, I. H., Waters, M., Grau, R. R., and Sarker, M. R. (2004). Disruption of the gene (*spoA*) encoding sporulation transcription factor blocks endospore formation and enterotoxin production in enterotoxigenic *Clostridium perfringens* type A. *FEMS Microbiol. Lett.* 233, 233–240. doi: 10.1111/j.1574-6968.2004.tb09487.x
- Ingram, M., and Robinson, R. H. M. (1951). A discussion of the literature on botulism in relation to acid foods. *Proc. Soc. Appl. Bacteriol.* 14, 73–84. doi: 10.1111/j.1365-2672.1951.tb01995.x
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031
- Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., et al. (2012). Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* 7, 1511–1522. doi: 10.1038/nprot.2012.085
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199
- Kenri, T., Sekizuka, T., Yamamoto, A., Iwaki, M., Komiya, T., Hatakeyama, T., et al. (2014). Genetic characterization and comparison of *Clostridium botulinum* isolates from botulism cases in Japan between 2006 and 2011. *Appl. Environ. Microbiol.* 80, 6954–6964. doi: 10.1128/AEM.02134-14
- Lee, K. S., Bumbaca, D., Kosman, J., Setlow, P., and Jedrzejas, M. J. (2008). Structure of a protein-DNA complex essential for DNA protection in spores of *Bacillus* species. *Proc. Natl. Acad. Sci. U.S.A.* 105, 2806–2811. doi: 10.1073/pnas.0708244105
- Li, J., and McClane, B. A. (2008). A novel small acid soluble protein variant is important for spore resistance of most *Clostridium perfringens* food poisoning isolates. *PLoS Pathog.* 4:e1000056. doi: 10.1371/journal.ppat.1000056
- Li, J., Paredes-Sabja, D., Sarker, M. R., and McClane, B. A. (2009). Further characterization of *Clostridium perfringens* small acid soluble protein-4 (Ssp4) properties and expression. *PLoS ONE* 4:e6249. doi: 10.1371/journal.pone.0006249
- Li, Y., Davis, A., Korza, G., Zhang, P., Setlow, B., Setlow, P., et al. (2012). Role of a SpoVA protein in dipicolinic acid uptake into developing spores of *Bacillus subtilis*. *J. Bacteriol.* 194, 1875–1884. doi: 10.1128/JB.00062-12
- McClung, L. S. (1937). Studies on anaerobic bacteria X. heat stable and heat labile antigens in the botulinus and related groups of sporebearing anaerobes. *J. Infect. Dis.* 60, 122–128. doi: 10.1093/infdis/60.1.122
- Molle, V., Fujita, M., Jensen, S. T., Eichenberger, P., González-Pastor, J. E., Liu, J. S., et al. (2003). The SpoA regulon of *Bacillus subtilis*. *Mol. Microbiol.* 50, 1683–1701. doi: 10.1046/j.1365-2958.2003.03818.x
- Nakamura, S., Okado, I., Nakashio, S., and Nishida, S. (1977). *Clostridium sporogenes* isolates and their relationship to *C. botulinum* based on deoxyribonucleic acid reassociation. *J. Gen. Microbiol.* 100, 395–401. doi: 10.1099/00221287-100-2-395
- Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x
- Onyenwoke, R. U., Brill, J. A., Farahi, K., and Wiegel, J. (2004). Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes). *Arch. Microbiol.* 182, 182–192. doi: 10.1007/s00203-004-0696-y
- Orsburn, B. C., Melville, S. B., and Popham, D. L. (2010). EtfA catalyses the formation of dipicolinic acid in *Clostridium perfringens*. *Mol. Microbiol.* 75, 178–186. doi: 10.1111/j.1365-2958.2009.06975.x
- Orsburn, B., Sucre, K., Popham, D. L., and Melville, S. B. (2009). The SpmA/B and DacF proteins of *Clostridium perfringens* play important roles in spore heat resistance. *FEMS Microbiol. Lett.* 291, 188–194. doi: 10.1111/j.1574-6968.2008.01454.x
- Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421
- Paidhungat, M., Setlow, B., Driks, A., and Setlow, P. (2000). Characterization of spores of *Bacillus subtilis* which lack dipicolinic acid. *J. Bacteriol.* 182, 5505–5512. doi: 10.1128/JB.182.19.5505-5512.2000
- Paredes-Sabja, D., Sarker, N., Setlow, B., Setlow, P., and Sarker, M. R. (2008a). Roles of DacB and Spm proteins in *Clostridium perfringens* spore resistance to moist heat, chemicals, and UV radiation. *Appl. Environ. Microbiol.* 74, 3730–3738. doi: 10.1128/AEM.00169-08
- Paredes-Sabja, D., Setlow, B., Setlow, P., and Sarker, M. R. (2008b). Characterization of *Clostridium perfringens* spores that lack SpoVA proteins and dipicolinic acid. *J. Bacteriol.* 190, 4648–4659. doi: 10.1128/JB.00325-08
- Perez-Valdespino, A., Li, Y., Setlow, B., Ghosh, S., Pan, D., Korza, G., et al. (2014). Function of the SpoVAE and SpoVAF proteins of *Bacillus subtilis* spores. *J. Bacteriol.* 196, 2077–2088. doi: 10.1128/JB.01546-14
- Popham, D. L., Gilmore, M. E., and Setlow, P. (1999). Roles of low-molecular-weight penicillin-binding proteins in *Bacillus subtilis* spore peptidoglycan synthesis and spore properties. *J. Bacteriol.* 181, 126–32.
- Ragkousi, K., Eichenberger, P., van Ooij, C., and Setlow, P. (2003). Identification of a new gene essential for germination of *Bacillus subtilis* spores with Ca²⁺-dipicolinate. *J. Bacteriol.* 185, 2315–2329. doi: 10.1128/JB.185.7.2315-2329.2003
- Raju, D., Waters, M., Setlow, P., and Sarker, M. R. (2006). Investigating the role of small, acid-soluble spore proteins (SASPs) in the resistance of *Clostridium perfringens* spores to heat. *BMC Microbiol.* 6:50. doi: 10.1186/1471-2180-6-50
- R Core Team T. (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rossetto, O., Pirazzini, M., and Montecucco, C. (2014). Botulinum neurotoxins: genetic, structural and mechanistic insights. *Nat. Rev. Microbiol.* 12, 535–549. doi: 10.1038/nrmicro3295
- Schill, K. M., Wang, Y., Butler, R. R. III., Pombert, J.-F., Reddy, N. R., Skinner, G. E., et al. (2016). Genetic diversity of *Clostridium sporogenes* PA 3679 isolates obtained from different sources as resolved by pulsed-field gel electrophoresis and high-throughput sequencing. *Appl. Environ. Microbiol.* 82, 384–393. doi: 10.1128/AEM.02616-15
- Setlow, P. (2006). Spores of *Bacillus subtilis*: their resistance to and killing by radiation, heat and chemicals. *J. Appl. Microbiol.* 101, 514–525. doi: 10.1111/j.1365-2672.2005.02736.x
- Setlow, P. (2007). I will survive: DNA protection in bacterial spores. *Trends Microbiol.* 15, 172–180. doi: 10.1016/j.tim.2007.02.004
- Setlow, P. (2014a). Germination of spores of *Bacillus* Species: what we know and do not know. *J. Bacteriol.* 196, 1297–1305. doi: 10.1128/JB.01455-13
- Setlow, P. (2014b). Spore Resistance Properties. *Microbiol. Spectr.* 2, 201–215. doi: 10.1128/microbiolspec.TBS-0003-2012
- Stumbo, C. R., Purohit, K. S., and Ramakrishnan, T. V. (1975). Thermal process lethality guide for low-acid foods in metal containers. *J. Food Sci.* 40, 1316–1323. doi: 10.1111/j.1365-2621.1975.tb01080.x
- Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039
- Tovar-Rojo, F., Chander, M., Setlow, B., and Setlow, P. (2002). The products of the *spoVA* operon are involved in dipicolinic acid uptake into developing spores of *Bacillus subtilis*. *J. Bacteriol.* 184, 584–587. doi: 10.1128/JB.184.2.584-587.2002

- Townsend, C. T., Esty, J. K., and Baselt, F. C. (1938). Heat-resistance studies on spores of putrefactive anaerobes in relation to determination of safe processes for canned foods. *J. Food Sci.* 3, 323–346. doi: 10.1111/j.1365-2621.1938.tb17065.x
- Velásquez, J., Schuurman-Wolters, G., Birkner, J. P., Abbe, T., and Poolman, B. (2014). *Bacillus subtilis* spore protein SpoVAC functions as a mechanosensitive channel. *Mol. Microbiol.* 92, 813–823. doi: 10.1111/mmi.12591
- Wang, S. T., Setlow, B., Conlon, E. M., Lyon, J. L., Imamura, D., Sato, T., et al. (2006). The forespore line of gene expression in *Bacillus subtilis*. *J. Mol. Biol.* 358, 16–37. doi: 10.1016/j.jmb.2006.01.059
- Weigand, M. R., Pena-Gonzalez, A., Shirey, T. B., Broeker, R. G., Ishaq, M. K., Konstantinidis, K. T., et al. (2015). Implications of genome-based discrimination between *Clostridium botulinum* group I and *Clostridium sporogenes* strains for bacterial taxonomy. *Appl. Environ. Microbiol.* 81, 5420–5429. doi: 10.1128/AEM.01159-15
- Wetzel, D., and Fischer, R.-J. (2015). Small acid-soluble spore proteins of *Clostridium acetobutylicum* are able to protect DNA *in vitro* and are specifically cleaved by germination protease GPR and spore protease YyaC. *Microbiology* 161, 2098–2109. doi: 10.1099/mic.0.000162
- Williamson, C. H., Sahl, J. W., Smith, T. J., Xie, G., Foley, B. T., Smith, L. A., et al. (2016). Comparative genomic analyses reveal broad diversity in botulinum-toxin-producing Clostridia. *BMC Genomics* 17:180. doi: 10.1186/s12864-016-2502-z

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Butler, Schill, Wang and Pombert. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genotypes Associated with *Listeria monocytogenes* Isolates Displaying Impaired or Enhanced Tolerances to Cold, Salt, Acid, or Desiccation Stress

Patricia Hingston¹, Jessica Chen^{1†}, Bhavjinder K. Dhillon², Chad Laing³, Claire Bertelli², Victor Gannon³, Taurai Tasara⁴, Kevin Allen¹, Fiona S. L. Brinkman², Lisbeth Truelstrup Hansen⁵ and Siyun Wang^{1*}

¹ Department of Food, Nutrition, and Health, University of British Columbia, Vancouver, BC, Canada, ² Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada, ³ Laboratory for Foodborne Zoonoses, Public Health Agency of Canada, Lethbridge, AB, Canada, ⁴ Institute for Food Safety and Hygiene, University of Zurich, Zurich, Switzerland, ⁵ Division for Microbiology and Production, National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark

OPEN ACCESS

Edited by:

Jennifer Ronholm,
McGill University, Canada

Reviewed by:

Laurent Guillier,
ANSES, France

David Rodriguez-Lazaro,
University of Burgos, Spain

*Correspondence:

Siyun Wang
siyun.wang@ubc.ca

†Present Address:

Jessica Chen,
IHRC Inc., Atlanta, GA, USA

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 16 December 2016

Accepted: 22 February 2017

Published: 08 March 2017

Citation:

Hingston P, Chen J, Dhillon BK, Laing C, Bertelli C, Gannon V, Tasara T, Allen K, Brinkman FSL, Truelstrup Hansen L and Wang S (2017) Genotypes Associated with *Listeria monocytogenes* Isolates Displaying Impaired or Enhanced Tolerances to Cold, Salt, Acid, or Desiccation Stress. *Front. Microbiol.* 8:369.

doi: 10.3389/fmicb.2017.00369

The human pathogen *Listeria monocytogenes* is a large concern in the food industry where its continuous detection in food products has caused a string of recalls in North America and Europe. Most recognized for its ability to grow in foods during refrigerated storage, *L. monocytogenes* can also tolerate several other food-related stresses with some strains possessing higher levels of tolerances than others. The objective of this study was to use a combination of phenotypic analyses and whole genome sequencing to elucidate potential relationships between *L. monocytogenes* genotypes and food-related stress tolerance phenotypes. To accomplish this, 166 *L. monocytogenes* isolates were sequenced and evaluated for their ability to grow in cold (4°C), salt (6% NaCl, 25°C), and acid (pH 5, 25°C) stress conditions as well as survive desiccation (33% RH, 20°C). The results revealed that the stress tolerance of *L. monocytogenes* is associated with serotype, clonal complex (CC), full length *inlA* profiles, and the presence of a plasmid which was identified in 55% of isolates. Isolates with full length *inlA* exhibited significantly ($p < 0.001$) enhanced cold tolerance relative to those harboring a premature stop codon (PMSC) in this gene. Similarly, isolates possessing a plasmid demonstrated significantly ($p = 0.013$) enhanced acid tolerance. We also identified nine new *L. monocytogenes* sequence types, a new *inlA* PMSC, and several connections between CCs and the presence/absence or variations of specific genetic elements. A whole genome single-nucleotide-variants phylogeny revealed sporadic distribution of tolerant isolates and closely related sensitive and tolerant isolates, highlighting that minor genetic differences can influence the stress tolerance of *L. monocytogenes*. Specifically, a number of cold and desiccation sensitive isolates contained PMSCs in σ^B regulator genes (*rsbS*, *rsbU*, *rsbV*). Collectively, the results suggest that knowing the sequence type of an isolate in addition to screening for the presence of full-length *inlA* and a plasmid, could help

food processors and food agency investigators determine why certain isolates might be persisting in a food processing environment. Additionally, increased sequencing of *L. monocytogenes* isolates in combination with stress tolerance profiling, will enhance the ability to identify genetic elements associated with higher risk strains.

Keywords: *Listeria monocytogenes*, stress tolerance, whole genome sequencing, sequence typing, food safety, plasmids, internalin A

INTRODUCTION

Listeria monocytogenes is a ubiquitous bacterial foodborne pathogen that is most recognized for its ability to grow at temperatures as low as -0.4°C (Walker et al., 1990) and cause listeriosis, a serious disease with an average mortality rate of 30% among at-risk people (Yildiz et al., 2007). In addition to possessing cold tolerance, *L. monocytogenes* is also capable of surviving many other food-related stresses including high osmolarity (Shabala et al., 2008) and low pH (Sorrells et al., 1989), further adding to its hardiness. Additionally, cross contamination of foods is facilitated by biofilm formation (Chavant et al., 2002; Mørerø and Langsrud, 2004; Moltz, 2005; Di Bonaventura et al., 2008; Hingston et al., 2013), and the ability of the organism to survive desiccation for extended periods of time on food contact surfaces (Vogel et al., 2010). Post-processing levels of *L. monocytogenes* contamination in foods are usually low (Fenlon et al., 1996; Kozak et al., 1996; Cabedo et al., 2008) and unlikely to cause disease (Buchanan et al., 1997; Chen et al., 2003). It is therefore, refrigerated, ready-to-eat (RTE) foods with extended shelf lives and the potential for regrowth that present the largest risk to consumers.

Both Canada and the EU have adopted regulations for the control of *L. monocytogenes* in RTE foods (Health Canada, 2011; Luber, 2011), allowing up to 100 CFU/g in foods that do not permit growth beyond this level within the shelf-life of the product, and a zero tolerance policy for foods identified as supporting growth. When validating growth inhibition of *L. monocytogenes* in stabilized RTE foods, it is important that the strains used represent the extremes of *L. monocytogenes'* stress response behavior. In the US, the zero tolerance policy is applicable for all food products (US FDA, 2016). However, nationwide outbreaks continue to occur in the US. To date, there have been three multistate listeriosis outbreaks in 2016 that were associated with frozen vegetables, packaged salads, and raw milk and resulted in 29 illnesses and 4 deaths (CDC, 2016a). These numbers are yet to exceed that of 2015 where two multistate outbreaks involving soft cheese and ice cream resulted in 40 illnesses and 6 deaths (CDC, 2016a).

In 2013, the US established the *Listeria* Whole Genome Sequencing (WGS) Project to assist in detecting, investigating, and mitigating foodborne outbreaks (CDC, 2016b). Though valuable for tracing outbreaks, WGS is not routinely used to determine the stress tolerance of outbreak strains. However, WGS provides the information that could potentially lead to identification of molecular biomarkers related to the stress tolerance of *L. monocytogenes* isolates. Such biomarkers could greatly aid in monitoring the risks of *L. monocytogenes*

contamination and regrowth in food products and processing environments (Jacquet et al., 2004).

Currently, one common molecular biomarker used for *L. monocytogenes'* virulence is the internalin A encoding gene (*inlA*) which can contain one of several different premature stop codons producing truncated and secreted proteins associated with attenuated virulence (Jonquieres et al., 1998; Jacquet et al., 2004; Rousseaux et al., 2004; Nightingale et al., 2005; Felicio et al., 2007; Handa-Miya et al., 2007; Roldgaard et al., 2009; Van Stelten et al., 2011). Recently, Kovacevic et al. (2013) discovered that full-length variants of *inlA* were more prevalent among fast cold-adapting *L. monocytogenes* strains than intermediate and slow cold-adapting strains, suggesting that *inlA* profiling may also be suitable for predicting the cold tolerance of strains. Another potential biomarker is the *L. monocytogenes* stress survival islet 1 (SSI-1). Included in this five gene cluster (*lmo0444-lmo0448*) are two genes (*gadT1* and *gadD1*) from the glutamate decarboxylase acid resistance system which has been shown to significantly improve the growth of *L. monocytogenes* in mildly acidic environments (Cotter et al., 2005). Additionally, an *L. monocytogenes* SSI-1 deletion mutant exhibited impaired growth at low pH (pH 4.8), high salt (7.5% NaCl), and on frankfurters stored at 4°C (Ryan et al., 2010). Further, research with naturally occurring SSI-1 positive and negative strains is needed to determine if this island would be a suitable biomarker for predicting stress tolerance phenotypes.

To date, studies which evaluated the stress tolerances of *L. monocytogenes* isolates have focused on associating phenotypes with genetic lineages (Bergholz et al., 2010), serotypes (Junttila et al., 1988; Barbosa et al., 1994; Ribeiro et al., 2014), and isolation sources (Begot et al., 1997; Durack et al., 2013). However, few significant differences between these groups were observed, suggesting that the diversity among isolates within these means of classification is not definitive for predicting phenotypic behavior. Instead, stronger phenotype associations might be observed among more closely related isolates, e.g., those sharing the same sequence type (ST) or clonal complex (CC). Additionally, the presence of specific genetic elements (e.g., *inlA* and SSI-1) may also influence the stress tolerance phenotypes of isolates as well as more minor genetic differences such as single nucleotide variants (SNVs).

The objective of this study was to use a combination of phenotypic analyses and WGS to elucidate novel associations between *L. monocytogenes* genotypes and food-related stress tolerance phenotypes with the goal of identifying biomarkers that can be used to predict the stress tolerances of food-chain isolates. To accomplish this, 166 *L. monocytogenes* isolates were evaluated on their ability to grow in cold (4°C), salt (6% NaCl),

and acid (pH 5) stress conditions as well as survive desiccation stress (33% RH). Factors investigated for potential associations with the observed phenotypes were: genetic lineage, serotype, CC, *inlA* profiles, and the presence of a plasmid, SSI-1, unique SNVs, and *Listeria* genomic island 1 (LGI1).

MATERIALS AND METHODS

Isolates and Culture Conditions

A collection of 166 *Listeria monocytogenes* isolates from Canada and Switzerland were used in this collaborative study. This included: (i) 159 food and food processing environment isolates from Canada ($n = 139$) and Switzerland ($n = 20$), (ii) six isolates from sporadic human listeriosis cases in Switzerland, and (iii) one isolate from an asymptomatic human (Table S1). All human isolates were anonymized and no ethical approval was required as per the institutional and national guidelines. Isolates were stored at -70°C in brain heart infusion broth (BHIB, Difco, Fisher Scientific, Canada) +20% glycerol and routinely cultured at 30°C on BHI agar (BHIA, Difco, Fisher Scientific) plates.

Whole Genome Sequencing

Genomic DNA was isolated using the PureLink Mini Kit from Life Technologies, Canada. PicoGreen quantification was performed (Invitrogen, Canada) and DNA was assessed using the NanoDrop 2000 (Fisher Scientific). Genomic DNA samples of sufficient quality and quantity were sequenced by Genome Quebec (Montréal, QC, Canada) using TruSeq automated library preparation (Illumina) and paired-end, 100 bp sequencing on the Illumina Hi-Seq. Between 4.9 and 16.5 million high quality reads remained after quality control for each genomic library. Raw FASTQ files were trimmed using Cutadapt in Trim Galore! version 0.4.1 and *de novo* genome assembly was performed using SPAdes version 3.1.0 (careful option used; Bankevich et al., 2012). Low coverage (<10) and small contigs (<200 bp) were removed from assemblies using a custom perl script. Assemblies were subsequently annotated using Prokka version 1.5.2 (genus *Listeria*, species *monocytogenes*; Seemann, 2014). Assembled sequences were deposited into the NCBI Whole Genome Shotgun (WGS) database under Bioproject PRJNA329415.

Lineage Determination

To classify isolates into genetic lineages, a reference free, k-mer based single nucleotide variants (SNV) phylogeny was generated using the kSNP 3.0 program (Gardner et al., 2015) and reference isolates for the major lineages of *L. monocytogenes* (LI—F2365; LII—EGD-e; LIII—HCC23; LIV—J1-208). The resulting maximum parsimony tree (based on the consensus of 100 trees) clearly segregated the four lineages.

Multi Locus Sequence Typing

To group isolates based on their epidemiological context, *in silico* MLST was performed using the Center for Genomic Epidemiology's MLST typing tool (<https://cge.cbs.dtu.dk/services/MLST/>). Clonal Complexes (CCs) were assigned based on the Pasteur Institute schema (<http://www.pasteur.fr/recherche/genopole/PF8/mlst/Lmono.html>). Novel sequence

types (STs) were confirmed using Sanger sequencing and submitted to the Pasteur Institute Database for new assignments (<http://bigsdb.pasteur.fr/listeria/listeria.html>).

In silico Serogroup/Serotype Assignment

Antibody-based serotyping was conducted on a subset of isolates ($n = 91$) within both the current study and previous studies (Arguedas-Villa et al., 2010; Kovacevic et al., 2013). Remaining isolates were assigned one or more possible serotypes by performing a MegaBLAST search (>95% nt identity) ncbi-blast+ v. 2.3.0 available at: <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>) for four genes used in a multiplex PCR developed by Doumith et al. (2004). Additionally, predictive serotypes were assigned to isolates with STs that are known to be associated with a specific serotype.

Targeted Genomic Element Screenings

The genes and genomic regions evaluated in this study were (1) the plasmid replicon gene *repA*, used to indicate the presence of a plasmid (Kuenne et al., 2010), (2) *emrE*, representing *Listeria* genomic island 1 (LGI1), a 50 kb island with putative roles in stress tolerance and persistence (Gilmour et al., 2010), and (3) stress survival islet 1 (SSI-1), a five gene cluster previously identified as having a role in *L. monocytogenes*' response to cold, osmotic, and acid stress conditions (Ryan et al., 2010). Additionally, the coding sequence of *inlA* was investigated to determine if isolates possessed a full length sequence or a premature stop codon (PMSC) mutation. *emrE*, SSI-1, and *inlA* were screened for among isolate sequence assemblies using MegaBLAST (>95% nt identity) and *repA* was screened for using BLASTP (>30% aa identity over >80% coverage). *inlA* and *repA* sequences were then extracted from the isolate assemblies for further analysis.

Identification of Putative Plasmid Contigs

Detection of *repA* sequences meant that at least one contig belonged to a putative plasmid. To identify additional plasmid associated contigs, isolate assemblies were aligned to the closed genome of *L. monocytogenes* EDG-e (Accession: NC_003210.1) using Contig Mover in Mauve version 2.3.1 (Rissman et al., 2009). Contigs not aligning to the EDG-e chromosome were compared to published *L. monocytogenes* plasmids (Kuenne et al., 2010) by BLAST. Contigs were excluded if they displayed open reading frames associated with chromosomal DNA (e.g., rRNA, tRNA) or did not align to any of the *Listeria* associated plasmids annotated by Kuenne et al. (2010): pLM33, pLM1-2bUG1, pLM5578, pLM80, and pLI100. A summary of the putative plasmid contigs found within each isolate can be found in Table S1.

Cold Tolerance Assay

Overnight cultures grown in BHIB at 37°C were standardized to 10^9 CFU/ml using spectrophotometric methods, and diluted in pre-chilled BHIB to yield a final density of 10^3 CFU/ml and stored at 4°C (previously described in Arguedas-Villa et al., 2010). The bacterial density was enumerated daily for the first 4 days and then bi-weekly for up to 5 weeks by plating on tryptic soy agar (BD, Fisher Scientific) +6% yeast extract (BD,

Fisher Scientific). The resulting growth curves were fitted using a four parameter logistic model described by Dalgaard and Koutsoumanis (2001).

Salt and Acid Tolerance Assay

Isolates were assessed for salt and acid tolerance using modified versions of published protocols (Cotter et al., 2005; Van Der Veen et al., 2008; Bergholz et al., 2010). In short, overnight cultures grown in BHIB at 30°C were diluted in either BHIB+6% (w/w) NaCl or BHIB adjusted to pH 5 (with 1 M HCl) to achieve a final concentration of 10^7 CFU/ml. From these cultures, 200 µl was added in duplicate (technical replicates) to 96-well-plates (Costar™ clear polystyrene, Fisher Scientific) that were incubated at 25°C in a microplate reader (Spectramax, V6.3; Molecular Devices, Sunnyvale, CA). A temperature of 25°C was used to assess isolate salt and acid tolerance under non-intracellular or cold stress conditions. The absorbance ($A_{600\text{nm}}$) of each well was recorded every 30 min until all isolates reached stationary phase (~26 h) and the resulting growth curves were fitted to the Baranyi and Roberts model (Baranyi and Roberts, 1994) using DMfit (v3.5) available on the ComBase browser (<http://browser.combase.cc/DMFit.aspx>).

Desiccation Tolerance Assay

Cultures grown for 24 h in BHIB at 20°C were diluted to 10^7 CFU/ml in buffered peptone water (BD, Fisher Scientific) and 10 µl (10^5 CFU) was spotted in duplicate (technical replicates) on the bottom of wells in lid-less 96-well-plates. The plates were then stored for 3 days at 20°C in desiccators (SICCO, Bohlender, Germany) pre-conditioned to 33% relative humidity (RH) using a saturated solution of MgCl₂ (protocol adapted from Hingston et al., 2015). The RH of the chambers was monitored throughout the desiccation periods using data loggers (included with desiccators). A temperature of 20°C was used to simulate desiccation conditions that might occur in a food plant. Following desiccation, the plates were rehydrated with 200 µl of BHIB, and incubated at 25°C in a plate reader where the $A_{600\text{nm}}$ of each well was recorded every 30 min until all isolates reached stationary phase (~24 h). The resulting growth curves were then fitted to the Baranyi and Roberts model and the model parameters recorded.

Phenotype Designations and Statistical Analyses

For all four stress exposure experiments, a minimum of two biological replicates with two technical replicates each, were conducted for all isolates. Based on the findings of Aryani et al. (2015), the data was standardized for biological variability between replicates by dividing isolate growth parameters by the median value for each experimental run, thereby making the median equal to 1. The median was selected for standardization rather than the mean to avoid the influence of very stress sensitive isolates. Model parameters (LPD, lag phase duration; μ_{max} , maximum growth rate; N_{max} , maximum cell density) were averaged across biological replicates and presented as standardized (std) values. For isolates where the average std values had a standard deviation (SD) >0.05 , additional replicates

were completed to obtain more representative means. Isolates were considered tolerant or sensitive to cold, salt, or acid stress if they had an average std- $\mu_{\text{max}} >$ or $<$ than 1 SD from the median, respectively. All remaining isolates were considered to have intermediate stress tolerance. For desiccation stress survival the model parameter of most interest was the LPD, indicating the time to (detectable) regrowth (TRG) that is negatively correlated with the number of cells, which survived the desiccation treatment. Isolates were classified as desiccation tolerant or sensitive if they had an average std-TRG $<$ or $>$ than 1 SD from the median, respectively. A standard curve generated using five cell levels (10^1 - 10^5 CFU) produced a correlation of $y = -0.25x + 2.07$ ($R^2 = 0.97$) where y is the TDR and x is the \log_{10} number of viable cells in each well following desiccation.

To elucidate potential associations between the factors we investigated, statistical tests were performed using IBM SPSS Statistics version 23. Specifically, individual two-tailed *T*-tests and one-way ANOVAs with Tukey *post-hoc* tests were used to compare the average standardized stress tolerance model parameters of two (\pm plasmid, \pm SSI-1, \pm LGI1, lineage I and II, *repA* group 1 vs. group 2 isolates) or more groups (serotypes, CCs, *inLA* profiles, and sensitive, intermediate and stress tolerant groups), respectively. G*Power 3.1.9.2 (Faul et al., 2007, 2009) was used to determine the minimum sample sizes required to ensure a power of 0.80 for all statistical tests. All data sets were accessed for outliers, homogeneity of variances (Levene's test), and normality (Shapiro-Wilk's test). Where homogeneity of variances could not be achieved, Welch *T*-tests and Welch ANOVAs in combination with Games-Howell *post-hoc* tests were used. *P*-values below 0.05 were considered significant for all comparisons.

Phylogenetic Reconstruction Based on Core Genome Single Nucleotide Variants

Parsnp, a tool within Harvest suite of tools (Treangen et al., 2014), was used to perform core genome alignment of all 166 *de novo* assembled genomes and the reference *L. monocytogenes* EGD-e strain in order to identify single nucleotide variants (SNVs) within the core genome. SNVs clustered within 20 base pairs were removed as these may indicate repetitive regions containing more erroneous SNV calls. The remaining high quality SNVs were used to generate maximum likelihood trees using the RaxML version 8 (Stamatakis, 2014) on the CIPRES science gateway (Miller et al., 2010) using default parameters (including the GTRCAT nucleotide model and 100 bootstrap replicates). Corresponding heatmaps containing additional genotype and phenotype information were generated in R version 2.15.1 (Team, 2016) using the heatmap.2 function from the gplots library.

SNV Detection

SNVs were also detected against the *Listeria monocytogenes* EGD-e (NC_003210.1) reference genome. SMALT version 0.7.6 (<http://www.sanger.ac.uk/science/tools/smalt-0>) with default parameters except “-i 330” was used to first align raw reads against the reference. Samtools version 1.2 (Li,

2011) was used on these assemblies to sort the aligned reads (“samtools sort”), remove potential PCR duplicates (“samtools rmdup”) and call the SNVs (“samtools mpileup”). Additional filtering of SNV calls included removing those with a read depth <50 and heterozygous genotypes (since our genomes are haploid) using the “bcftools filter” command. SNVs found in repetitive regions of the genome as assessed by the index of repetitiveness (Schwender et al., 2004) were also removed manually. The remaining high confidence SNVs were annotated using SNPEff version 4.1 (Cingolani et al., 2012) with the *Listeria_monocytogenes_EGD_e_uid61583* annotation. Synonymous SNVs were also removed in the end for identification of non-synonymous or potential regulatory SNVs that may be contributing to phenotypic differences in cold growth.

Statistical Methods for Elucidating SNVs Associated with Stress Tolerance Phenotypes

SNPSift version 4.1 “CaseControl” (Cingolani et al., 2012) was used to run a Fisher Exact test to identify SNVs that were significantly associated with case vs. control groups. To identify SNVs only found in tolerant isolates, these were used as the case group, while all others were used as the control group. Since this did not yield any results, subsequently, the sensitive isolates alone were used as the control group so as to allow SNVs to be seen in intermediate growers. This method has certain limitations in that certain associations may require very large sample sizes to become statistically significant, especially considering genetic heterogeneity leads to the same phenotype or the potential for multiple SNVs to interact. An alternative approach, Random Forests™ (Breiman, 2001), was also used to discover important SNVs in distinguishing stress tolerant and sensitive groups since previous genome wide association studies have shown random forests outperform the Fisher Exact test in these special cases (Lunetta et al., 2004; Schwender et al., 2004; Bureau et al., 2005; De Lobel et al., 2010; Bulinski et al., 2011). The RandomForest™ version 4.6-10 library was used in R with default parameters except “importance = TRUE, proximity = TRUE, ntree = 5000.” This allows the method to run as a classifier that then ranks SNVs on their ability to classify isolates based on their phenotypic designation.

Genomic Islands Analysis

Annotated draft genomes were submitted to IslandViewer 3 (Dhillon et al., 2015) using *L. monocytogenes* EGD-e (NC_003210.1) as a reference for contig reordering. Genomic islands were predicted using IslandPath-DIMOB (Hsiao et al., 2005) and SIGI-HMM (Waack et al., 2006). Predicted genomic islands positioned on the genome within <10 kb of each other were merged into one single region.

To form groups of similar genomic islands, the genetic distance between genomic island sequences was computing using Mash (parameter -s 2000; Ondov et al., 2016) and groups of similar sequences were identified using hclust and cutree in R.

RESULTS

Genetic Characteristics of *L. monocytogenes* Isolates Based on WGS Data

The complete sequenced genome assembly sizes of the isolates ranged from 2.56 to 3.13 Mbp with a mean size of 2.97 Mbp (Table S1). Isolates belonged to one of three different lineages: LI ($n = 44$, serotypes 4b, 1/2b, 3b, and 3c), LII ($n = 121$, serotypes 1/2a, 1/2c, and 3a), and LIII ($n = 1$, serotype 4c). The majority of isolates were serotype 1/2a ($n = 92$), followed by 1/2c and 4b ($n = 25$ each), 1/2b ($n = 18$), 3a ($n = 2$), and 3b and 4c ($n = 1$ each; Table 1). The exact serotype was not determined for two remaining isolates. Beyond serotypes, our isolates belonged to 36 different known STs and a further nine were assigned novel STs (ST1017-1025). Isolates also belonged to one of 29 different CCs with a further seven isolates being unique non-clonal singletons. The most prevalent CCs in the collection were CCs 9, 8, and 7 (Table 2). Other less common CCs in decreasing prevalence included CCs 11, 155, 1, 3, and 321 (Table 2). Interestingly, only one CC121 isolate existed in our collection. This is surprising given that CC121 is often highly prevalent among *L. monocytogenes* food-associated isolates (Parisi et al., 2010; Chenal-Francisque et al., 2011; Martín et al., 2014; Ebner et al., 2015; Maury et al., 2016).

The plasmid replication gene, *repA*, was observed in 55% ($n = 92$) of our isolates, with a prevalence of 41 and 61% among LI and LII isolates, respectively. Notably, one isolate was observed to contain two putative plasmids as indicated by the presence of two different *repA* containing contigs of 61 and 69 kb. Among serotypes, *repA* was present in 100% of 3a isolates, 84% of 1/2c isolates, 78% of 1/2b isolates, 53% of 1/2a isolates, and 16% of 4b isolates (Table 1). Among CCs, plasmids were observed in >80% of CC 3, 5, 9, 11, and 321 isolates (Table 2).

A phylogeny, constructed on *repA* sequences as described in Kuenne et al. (2010), divided the sequences into two groups. Group 1 included isolates from serotypes 1/2a, 1/2b, 1/2c, and 4b with estimated plasmid sizes ranging from 26 to 88 kb. Group 2 included 1/2a, 1/2b, 1/2c, and 3a serotype isolates, harboring significantly larger plasmids ($p < 0.0005$, 55–100 kb) than those from group 1. These sizes are in line with those observed in Kuenne et al. (2010), supporting the assertion that these contigs belong to plasmids. The most prevalent plasmid size (56553–56554 bp) was observed for 26 isolates from seven different CCs and from both lineages I and II. Also noteworthy is that isolates from Switzerland and Canada contained plasmids with 100% nucleotide identity.

Premature stop codons (PMSCs) in *inlA* were observed in 20% of our isolates encompassing seven (Table S1) of 19 published PMSCs (Jonquieres et al., 1998; Olier et al., 2002, 2003; Rousseaux et al., 2004; Nightingale et al., 2005, 2008; Orsi et al., 2007; Van Stelten and Nightingale, 2008; Van Stelten et al., 2010; Wu et al., 2016) and one novel PMSC at 760aa, which was identified in two serotype 1/2c isolates. The most common PMSC occurred at 9aa ($n = 13$) and was associated with CC9, serotype 1/2c isolates. Ten of the 14 remaining CC9 isolates also had one of four *inlA* PMSCs (326, 576, 685, 760aa) and all CC321 isolates contained

TABLE 1 | Genetic characteristics and prevalence of sensitive and tolerant phenotypes among *L. monocytogenes* belonging to different serotypes.

Serotype	n (%)	Plasmid+ (%) ^a	Full length <i>inlA</i> (%) ^a	SSI-1+ (%) ^a	CS (%) ^a	CT (%) ^a	SS (%) ^a	ST (%) ^a	AS (%) ^a	AT (%) ^a	DS (%) ^a	DT (%) ^a
4b	25 (15)	4 (16)	14 (56)	4 (16)	4 (16)	2 (8)	1 (4)	3 (12)	1 (4)	4 (16)	2 (8)	3 (12)
1/2b	18 (11)	14 (78)	15 (83)	17 (94)	0	2 (11)	3 (17)	2 (11)	0	7 (39)	2 (11)	4 (22)
1/2a	92 (55)	49 (53)	85 (92)	65 (71)	4 (4)	11 (12)	18 (20)	12 (13)	22 (24)	7 (8)	10 (11)	11 (12)
1/2c	25 (15)	21 (84)	3 (12)	25 (100)	5 (20)	3 (12)	4 (16)	0	2 (8)	4 (16)	4 (16)	4 (16)
3a	2 (1)	2 (100)	0	2 (100)	0	0	1	0	1	0	0	0
3b	1	1	1	1	0	0	0	0	0	0	0	1
4c	1	0	0	0	0	0	0	0	0	0	1	0
1/2b, 3b, 7	1	0	1	1	0	0	0	0	0	0	0	0
1/2a, 3a	1	1	0	1	0	0	1	0	0	0	1	0
Sum	166	92	119	116	13	18	27	17	26	22	20	23

CS, cold sensitive; CT, cold tolerant; SS, salt sensitive; ST, salt tolerant; AS, acid sensitive; AT, acid tolerant; DS, desiccation sensitive; DT, desiccation tolerant.

^aPercentages relate to prevalence within the serotype.

inlA PMSC's at 700aa (Table S1). An additional 13 isolates, all from the serotype 4b CCs 6 and 315, contained a three codon deletion mutation previously reported in Kovacevic et al. (2013). With the exception of CCs 5 and 9, all isolates from the same CC either contained full length *inlA* or a truncated version.

During the screening of the whole genome sequences, the absence of *lmo1078* was noted among serotype 4b isolates. This gene, which encodes a UDP-glucose pyrophosphorylase, has been previously demonstrated to have a role in *L. monocytogenes* cold growth (Chassaing and Auvray, 2007). It was also observed that 70% (*n* = 116) of strains possessed SSI-1 with this island being most prevalent among serotype 1/2c isolates (100%) followed by 1/2b (94%), 1/2a (71%), and 4b (16%; Table 2). Furthermore, all isolates from CCs 3, 5, 7, 8, 9, 155, 224, 315, and 321 contained SSI-1 (Table 2). All remaining isolates possessed a homolog to F2365_0481 in place of SSI-1 (Ryan et al., 2010), with the exception of the CC121 isolate which possessed *lin0464* and *lin0465* homologs as reported in Hein et al. (2011).

The LGI1 indicator gene, *emrE*, was found in 16 of our isolates and as previously reported (Gilmour et al., 2010; Althaus et al., 2014; Kovacevic et al., 2015), all originated from Canada and 14 were serotype 1/2a ST120-CC8. The remaining two isolates represented novel STs (ST1022 and 1025) that also belonged to CC8. All *emrE* containing isolates also harbored SSI-1 and full length *inlA*.

Stress Tolerance Distributions among *L. monocytogenes* Isolates

All *L. monocytogenes* isolates were evaluated on their ability to grow in cold (4°C), salt (6% NaCl), and acid (pH 5) stress conditions as well as survive desiccation stress (33% RH). The cold growth plate count data was modeled using the Dalgaard and Koutsoumanis (2001) logistic model because it was more accommodating of fewer sampling points [average R^2 = 0.998, mean standard error (MSE) = 0.129]. From the std- μ_{\max} values, 13 isolates were classified as cold sensitive and 18 were classified as cold tolerant with average std- μ_{\max} values of 0.85 ± 0.08 and 1.09 ± 0.02 , respectively (Figure 1A). For the salt, acid, and

desiccation tolerance assays, the Baranyi and Roberts (1994) was suitable for modeling the spectrophotometrically obtained data with average R^2 and MSE-values ranging from 0.997 to 0.998 and 0.003 to 0.017, respectively. Overall, 27 and 17 isolates were classified as salt sensitive and tolerant with average std- μ_{\max} values of 0.83 ± 0.05 and 1.16 ± 0.05 (Figure 1B); 26 and 22 isolates were classified as acid sensitive and tolerant with average std- μ_{\max} values of 0.64 ± 0.14 and 1.34 ± 0.12 (Figure 1C); and 20 and 23 isolates were identified as desiccation sensitive and tolerant isolates with average std-TRGs of 0.81 ± 0.06 and 1.22 ± 0.11 (Figure 1D), respectively.

Overlapping Stress-Tolerance Phenotypes

Five isolates were classified as sensitive to three out of four stresses and 10 isolates were sensitive to two stresses, seven of which were salt and acid sensitive (Figure 2A). Only two isolates were classified as tolerant to three out of the four stresses and another 14 isolates were tolerant to two of the four stresses (Figure 2B). Twenty additional isolates displayed a total of 16 combinations of overlapping sensitive and tolerant phenotypes (Table S1). The most common overlapping phenotypes were salt and acid sensitive (*n* = 10), salt sensitive and desiccation tolerant (*n* = 6), cold and acid tolerant (*n* = 5), and cold tolerant and salt sensitive (*n* = 5).

Isolates designated sensitive, intermediate or tolerant to one stress were analyzed to determine if they significantly differed in their tolerances to other stresses. When grown in 6% NaCl, acid tolerant isolates had larger std- μ_{\max} values compared to acid sensitive isolates ($p = 0.03$, $\bar{x} = 1.02$ vs. 0.95). Similarly, salt sensitive isolates had smaller std- μ_{\max} values ($\bar{x} = 0.80$) in BHIB pH 5 compared to intermediate ($p < 0.0005$, $\bar{x} = 1.03$) and salt tolerant isolates ($p = 0.006$, $\bar{x} = 1.00$).

Stress Tolerances of *L. monocytogenes* Lineages, Serotypes, and Clonal Complexes

Between lineages, the only significant difference observed was that LI isolates had significantly larger std- μ_{\max} values in

TABLE 2 | Genetic characteristics and prevalence of sensitive and tolerant phenotypes among *L. monocytogenes* belonging to different clonal complexes.

CC	Associated serotypes	n (%)	Plasmid+ (%) ^a	Full length <i>imA</i> (%) ^a	SSI-1+ (%) ^a	CS (%) ^a	CT (%) ^a	SS (%) ^a	ST (%) ^a	AS (%) ^a	AT (%) ^a	DS (%) ^a	DT (%) ^a
1	4b	6 (4)	0	6 (100)	0	2 (33)	0	1 (17)	0	0	1 (17)	1 (17)	0
2	4b	3 (2)	0	3 (100)	0	1 (33)	0	0	2 (67)	0	0	0	0
3	1/2b	6 (4)	5 (83)	6 (100)	6 (100)	0	0	1 (17)	0	0	1 (17)	1 (17)	1 (17)
4	4b	3 (2)	0	3 (100)	0	0	1 (33)	0	0	0	0	2 (67)	1 (33)
5	1/2b	7 (4)	7 (100)	4 (57)	7 (100)	0	1 (14)	0	1 (14)	0	0	4 (57)	1 (14)
6	4b	7 (4)	4 (57)	0	0	1 (14)	0	0	0	1 (14)	0	0	2 (29)
7	1/2a	17 (10)	10 (59)	17 (100)	17 (100)	2 (12)	1 (6)	8 (47)	0	5 (29)	0	1 (6)	3 (18)
8	1/2a	22 (13)	14 (64)	22 (100)	22 (100)	0	3 (14)	2 (9)	1 (5)	1 (5)	3 (14)	2 (9)	0
9	1/2c, 1/2a	27 (16)	23 (85)	4 (15)	27 (100)	5 (19)	4 (15)	6 (22)	0	3 (11)	4 (15)	6 (22)	4 (15)
11	1/2a	11 (7)	10 (91)	11 (100)	0	0	1 (9)	0	4 (36)	0	1 (9)	0	4 (36)
14	1/2a	3 (2)	0	3 (100)	0	0	1 (33)	0	1 (33)	0	0	0	0
20	1/2a	3 (2)	0	3 (100)	0	0	1 (33)	1 (33)	1 (33)	2 (67)	0	0	0
155	1/2a	8 (5)	1 (13)	8 (100)	8 (100)	1 (13)	0	2 (25)	1 (13)	3 (38)	1 (13)	1 (13)	0
224	1/2b	4 (2)	0	4 (100)	4 (100)	0	1 (25)	2 (50)	0	0	1 (25)	0	3 (75)
315	4b	4 (2)	0	4 (100)	0	0	0	0	0	0	0	0	1 (25)
321	1/2a, 3a	6 (4)	6 (100)	6 (100)	6 (100)	0	0	1 (17)	5 (83)	0	0	2 (33)	0
Other	1/2a, 1/2b, 4b, 4c	29 (17)	12 (41)	25 (86)	15 (52)	1 (3)	4 (14)	4 (14)	5 (17)	6 (20)	4 (14)	6 (20)	2 (7)
Sum		166	92	119	116	13	18	27	17	26	22	20	23

CS, cold sensitive; CT, cold tolerant; SS, salt sensitive; ST, salt tolerant; AS, acid sensitive; AT, acid tolerant; DS, desiccation sensitive; DT, desiccation tolerant.

^aPercentages relate to prevalence within CCs.

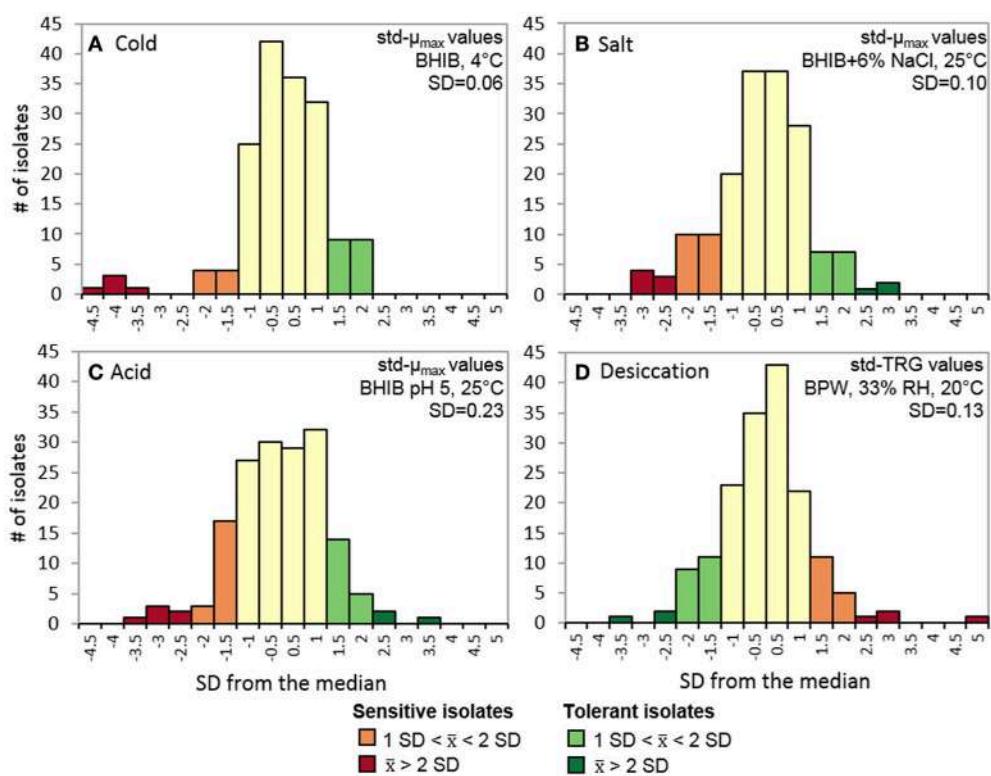


FIGURE 1 | Stress tolerance distributions of 166 *L. monocytogenes* isolates. Std- μ_{max} of isolates grown in (A) BHIB at 4°C, (B) BHIB+6% NaCl at 25°C, and (C) BHIB pH 5 at 25°C. (D) std-TRG of isolates after being desiccated at 33% RH for 3 days in BPW at 20°C and then rehydrated with BHIB and grown at 30°C. Isolates were classified as sensitive or tolerant if they displayed an average std- μ_{max} or std-TRG > 1 SD from the median (=1). std- μ_{max} , standardized maximum growth rate; std-TRG, standardized time to detectable regrowth; BHIB, brain heart infusion broth; BPW, buffered peptone water.

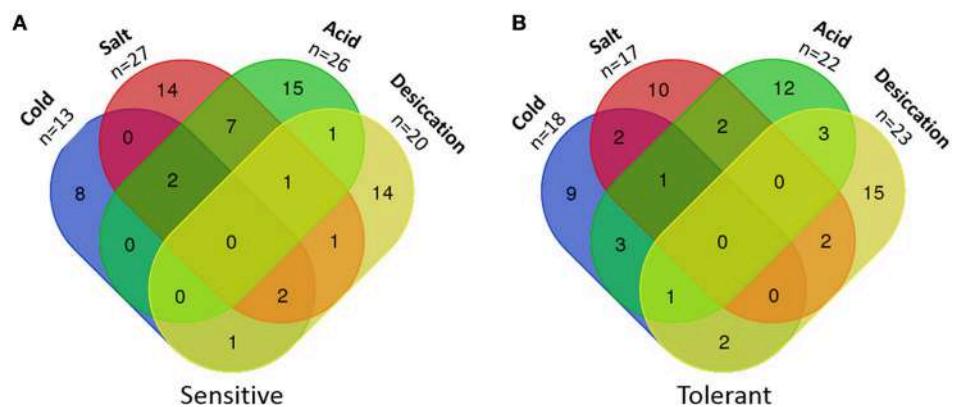


FIGURE 2 | Numbers of *L. monocytogenes* isolates with multiple sensitivities or tolerances to food-related stresses. (A) Sensitive isolates. (B) Tolerant isolates.

BHIB pH 5 than LII isolates ($p < 0.0005$, $\bar{x} = 1.13$ vs. $\bar{x} = 0.94$). Additional significant differences were observed between serotypes. At 4°C, serotype 1/2a isolates had significantly larger ($p = 0.017$) std- μ_{max} values compared to serotype 1/2c isolates (Figure 3). In support of this, serotype 1/2a isolates accounted for 61% of the cold tolerant isolates and only 31%

of cold sensitive isolates compared to a 55% prevalence of this serotype in the collection (Table 1). Similarly, serotype 1/2c isolates accounted for 38% of cold sensitive isolates despite a 15% overall prevalence in the collection (Table 1). When isolates were grown in 6% NaCl, no significant differences were observed between serotypes (Figure 3), however, 71% of salt tolerant and

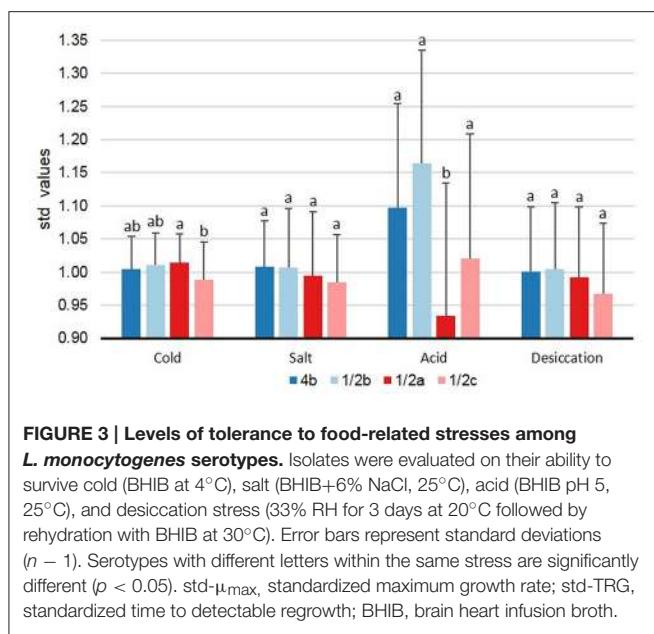


FIGURE 3 | Levels of tolerance to food-related stresses among *L. monocytogenes* serotypes. Isolates were evaluated on their ability to survive cold (BHIB at 4°C), salt (BHIB+6% NaCl, 25°C), acid (BHIB pH 5, 25°C), and desiccation stress (33% RH for 3 days at 20°C followed by rehydration with BHIB at 30°C). Error bars represent standard deviations ($n = 1$). Serotypes with different letters within the same stress are significantly different ($p < 0.05$). std- μ_{max} , standardized maximum growth rate; std-TRG, standardized time to detectable regrowth; BHIB, brain heart infusion broth.

67% of salt sensitive isolates were serotype 1/2a isolates, relative again to a prevalence of 55% in the collection (Table 1).

In BHIB pH 5, serotype 1/2a isolates had significantly smaller ($p = 0.027$) std- μ_{max} values than serotypes 1/2b, 1/2c, and 4b (Figure 3). In agreement with these findings, 85% of acid sensitive isolates were serotype 1/2a whereas only 32% of acid tolerant isolates were serotype 1/2a (Table 1). No significant differences ($p > 0.05$) were observed between serotypes with respect to desiccation stress std-TRGs (Figure 3).

Beyond the stress tolerances of lineages and serotypes, some significant differences were also detected between CCs. As a minimum of six isolates per CC were needed to ensure a power >0.80 for ANOVA results, statistical analyses were only performed using CCs 1, 3, 5, 6, 7, 8, 9, 11, 155, and 321. Figure 4 shows the average levels of cold (std- μ_{max}), salt (std- μ_{max}), acid (std- μ_{max}), and desiccation (std-TRG) tolerance among CCs with three or more isolates. At 4°C, no significant differences were found between the growth rates of different CCs, however, it was interestingly to see that CCs associated with 4b isolates had both the lowest and highest average average std- μ_{max} values at 4°C, demonstrating why stress tolerance differences were not observed between this serotype and others at 4°C (Figure 4A).

In 6% NaCl, CC7 (1/2a) isolates had significantly ($p < 0.05$) smaller std- μ_{max} values compared to CCs 5 (1/2b), 8 (1/2a), 11 (1/2a), and 155 (1/2a; Figure 4B). This highlights the range of salt tolerances between CCs within the same serotype and again explains why no significant differences were observed at the serotype level for salt tolerance. In support of the results shown in Figure 4B, 67% of CC2 isolates were salt tolerant while 50% of CC224, 47% of CC7, and 22% of CC9 isolates were salt sensitive (Table 2).

In BHIB pH 5, CC5 (1/2b) isolates exhibited significantly ($p < 0.05$) larger std- μ_{max} values than CCs 7, 155, and 321, and CC321 isolates additionally had smaller ($p < 0.05$) std- μ_{max}

values compared to CCs 1, 3, and 11 (Figure 4C). CC1 (4b) isolates also had significantly ($p = 0.02$) smaller std- μ_{max} values compared to CC7 (1/2a) isolates. From Figure 4C it can be seen that lineage I isolates were more acid tolerant than LII isolates, as the five CCs with the highest average std- μ_{max} values in BHIB pH 5 were all from LI while the five CCs with the lowest average std- μ_{max} values belonged to LII (predominantly 1/2a isolates). Notably, 57% of CC5 and 67% of CC4 isolates were acid tolerant while 83% of CC321, 67% of CC20, and 29% of CC7 isolates were acid sensitive (Table 2).

No significant differences were found between the desiccation stress std-TRGs of different CCs (Figure 4D). Nevertheless, CC224 (1/2b) had the smallest average std-TRGs and correspondingly, 75% of these isolates were classified as desiccation tolerant. CC11 (1/2a) had the next smallest std-TRGs while CCs 1 and 4 (both 4b) had the two largest average std-TRGs.

Associations between Plasmid Harborage and Stress Tolerances

Although plasmids were identified in 55% of all isolates, a higher percentage of plasmid carriers were observed among acid tolerant (73%), desiccation sensitive (75%), and desiccation tolerant (60%) isolates as compared to cold tolerant (33%) and acid sensitive (46%) isolates (Figure 5). Within LII, plasmid-positive isolates had smaller std- μ_{max} values at 4°C ($p = 0.024$, $\bar{x} = 1.00$ vs. 1.02) and larger std- μ_{max} ($p < 0.0005$, $\bar{x} = 1.01$ vs. 0.86) values when grown in BHIB pH 5 compared to their plasmid-free counterparts. No significant differences were found between the stress tolerance levels of LI plasmid-harboring and plasmid-free isolates. It was, however, observed that isolates containing *repA* group 1 plasmids; which were significantly ($p < 0.0005$) smaller than group 2 plasmids, had smaller std- μ_{max} ($p = 0.002$, $\bar{x} = 0.98$ vs. 1.04) values in 6% NaCl.

Associations between *inLA* Profiles and Stress Tolerances

Full length *inLA* was observed in 72% of isolates, where a higher percentage of the intact gene prevailed among cold (89%), salt (94%), and acid (82%) tolerant isolates while a lower prevalence was detected among desiccation tolerant isolates (35%; Figure 5). Statistically, isolates with full length *inLA* had significantly larger std- μ_{max} values at 4°C than isolates with an *inLA* PMSC ($p = 0.001$, $\bar{x} = 1.01$ vs. 0.97). Additionally, serotype 4b isolates possessing a three-codon deletion in *inLA*, had significantly shorter desiccation stress std-TRGs ($p = 0.002$, $\bar{x} = 0.94$ vs. 1.05) compared to serotype 4b isolates with full length *inLA*. No significant associations were found between *inLA* profiles and salt or acid stress tolerance.

Associations between Stress Tolerances and the Presence of SSI1 or LGI1

In 6% NaCl, isolates containing SSI-1 had significantly smaller std- μ_{max} values ($p = 0.004$, $\bar{x} = 0.98$ vs. 1.03) than isolates without SSI-1 though this difference was not large. To determine potential associations between LGI1 harborage and

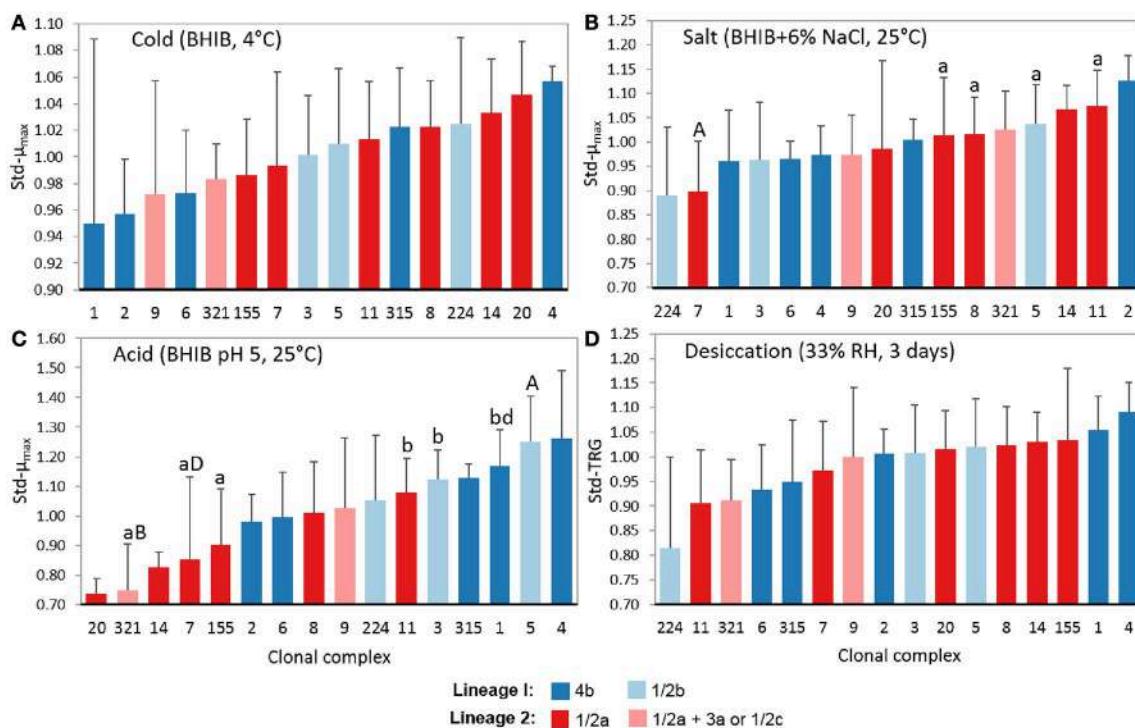


FIGURE 4 | Levels of tolerance to food-related stresses of different *L. monocytogenes* clonal complexes. Isolates were evaluated on their ability to survive (A) cold (BHIB at 4°C), (B) salt (BHIB+6% NaCl, 25°C), (C) acid (BHIB pH 5, 25°C), and (D) desiccation stress (33% RH for 3 days at 20°C followed by rehydration with BHIB at 30°C). Error bars represent standard deviations ($n = 1$) of standardized model values. CCs with different cases of the same letter are significantly different ($p < 0.05$). std- μ_{\max} , standardized maximum growth rate; std-TRG, standardized time to detectable regrowth; BHIB, brain heart infusion broth.

stress tolerance, serotype 1/2a isolates containing the island were compared to other 1/2a LGI1-negative isolates but similar to SSI-1, no significant differences in stress tolerances were detected.

SNV Analyses of Stress-Sensitive and Tolerant Isolates

Figure 6 shows a whole genome SNV phylogeny of all 166 *L. monocytogenes* isolates with their corresponding genetic and phenotypic properties. In this figure, groups of closely related isolates that share the same phenotypes can be seen. However, also shown are several cases where neighboring isolates have opposing stress tolerances. Of particular interest was whether specific SNVs could be related to isolates possessing the same stress tolerance phenotypes, however, none were detected to be uniquely shared among stress tolerant isolates that weren't also seen in intermediate or sensitive isolates. Among stress sensitive isolates, unique SNVs shared by subsets of isolates were identified, but no single SNV was prevalent among >4 isolates from the same stress sensitive phenotype group. In contrast, a large number of SNVs were uniquely observed for one or two isolates from the same stress sensitive group, causing frameshifts, premature stop codons, loss of start codons, or missense variants. Information regarding the SNVs identified among all sensitive and tolerant isolates are presented in Tables S2–S9. Notably, a number of stress sensitive isolates contained different PMSCs

in several σ^B regulator genes. A cold and desiccation sensitive isolate contained a PMSC in *rsbS* as did two other desiccation sensitive isolates. Furthermore, an additional cold sensitive isolate contained a PMSC in *rsbV*, and two desiccation sensitive isolates contained PMSCs in *rsbU*.

Genomic Islands of Stress-Sensitive and Tolerant Isolates

All *L. monocytogenes* isolates were predicted to harbor 1,318 genomic islands in total, resulting in an average of eight genomic islands per genome. These islands were clustered into 200 groups of similar sequences. The conservation of genomic island groups across *L. monocytogenes* lineages ranged from unique to a single isolate to conserved in 97 isolates (58%). Most frequently, intermediate groups conserved in subsets of monophyletic isolates were observed, as can be expected from a combination of vertical inheritance of the genomic islands with further modifications by mutation, insertions, and deletions. The clustering of *L. monocytogenes* isolates based on the presence or absence of groups of genomic islands reflected their phylogenetic proximity and did not relate particular genomic island content in isolates with the exhibition of similar phenotypes. Indeed, no single genomic island was found to occur in a large proportion of strains with a given phenotype (Figure S1).

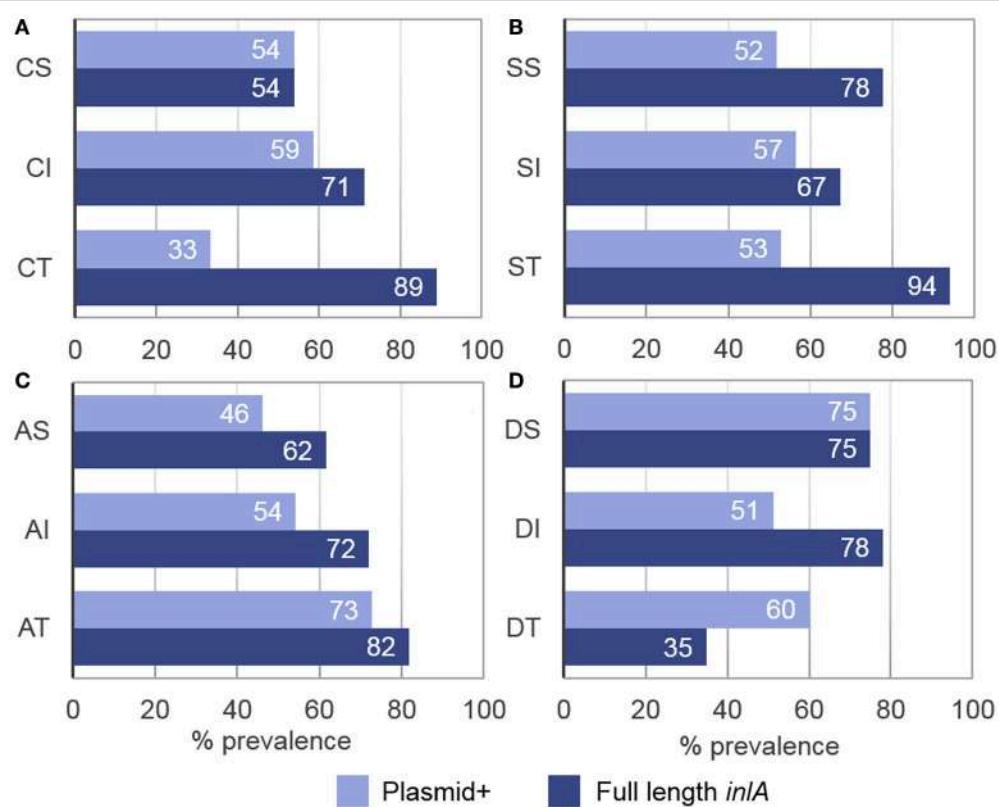


FIGURE 5 | Prevalence of full length *inlA* and plasmid harborage among *L. monocytogenes* stress tolerance phenotypes. **(A)** Cold sensitive (CS), intermediate (CI), and tolerant (CT) isolates. **(B)** Salt sensitive (SS), intermediate (SI), and tolerant (ST) isolates. **(C)** Acid sensitive (AS), intermediate (AI), and tolerant isolates (AT). **(D)** Desiccation sensitive (DS), intermediate (DI), and tolerant (DT) isolates. Full length *inlA* and the presence of a plasmid were observed in 72 and 55% of all isolates, respectively.

DISCUSSION

L. monocytogenes' Tolerance to Food-Related Stresses Differs between and within Lineages, Serotypes, and Clonal Complexes

Cold Stress

L. monocytogenes' ability to grow at refrigeration temperatures highlights this pathogen as a concern for the food industry and consumers alike. However, it is known that *L. monocytogenes* strains can largely differ in their ability to adapt to cold stress. In the present study, it was found that serotypes 1/2a and 1/2b were on average more cold-tolerant than serotypes 4b and 1/2c. Other cold growth studies have also reported serotype 1/2a strains to be more cold tolerant than serotype 4b strains (Junttila et al., 1988; Buncic et al., 2001; Lianou et al., 2006) though similar to the current findings, many of these differences were not statistically significant due to strain to strain variations. Barbosa et al. (1994) reported that out of 39 *L. monocytogenes* strains, Scott A, a 4b clinical isolate, grew the slowest at 4°C and that 1/2a strains grew the fastest followed by 1/2b, and 4b. Similarly, in De Jesús and Whiting (2003), LII isolates (all serotype 1/2a) exhibited the shortest LPDs at 5°C followed by LI

and then LIII isolates. Researchers have suggested that LII strains may be able to survive better under food-related stresses due to an enhanced ability to acquire advantageous mutations and extrachromosomal DNA compared to LI strains which typically have more conserved genomes (Orsi et al., 2007, 2008, 2011; Ragon et al., 2008; Dunn et al., 2009). Certain stress response genes, predominantly involved in membrane transport and cell wall structure (Doumith et al., 2004), have also been reported to be present in LII isolates but absent among LI isolates (Borucki and Call, 2003; Call et al., 2003; Zhang et al., 2003; Doumith et al., 2004; Chan and Wiedmann, 2008). Given the critical roles of these structures in allowing bacteria to adapt and tolerate numerous stresses (Annous et al., 1997; Verheul et al., 1997; Klein et al., 1999; Weber et al., 2001; Álvarez-Ordóñez et al., 2008), it is not surprising that *L. monocytogenes* lineages and serotypes can behave differently under certain stresses. The alternative sigma factor, σ^C , and *lmo1078*, encoding a UDP-glucose pyrophosphorylase, are examples of genes with reported roles in *L. monocytogenes* cold tolerance, that are present in LII strains but absent in LI and serotype 4b strains, respectively (Chassaing and Auvray, 2007; Chan and Wiedmann, 2008). These absences may partly explain the overall reduced cold tolerance of serotype 4b isolates respective to 1/2a isolates.

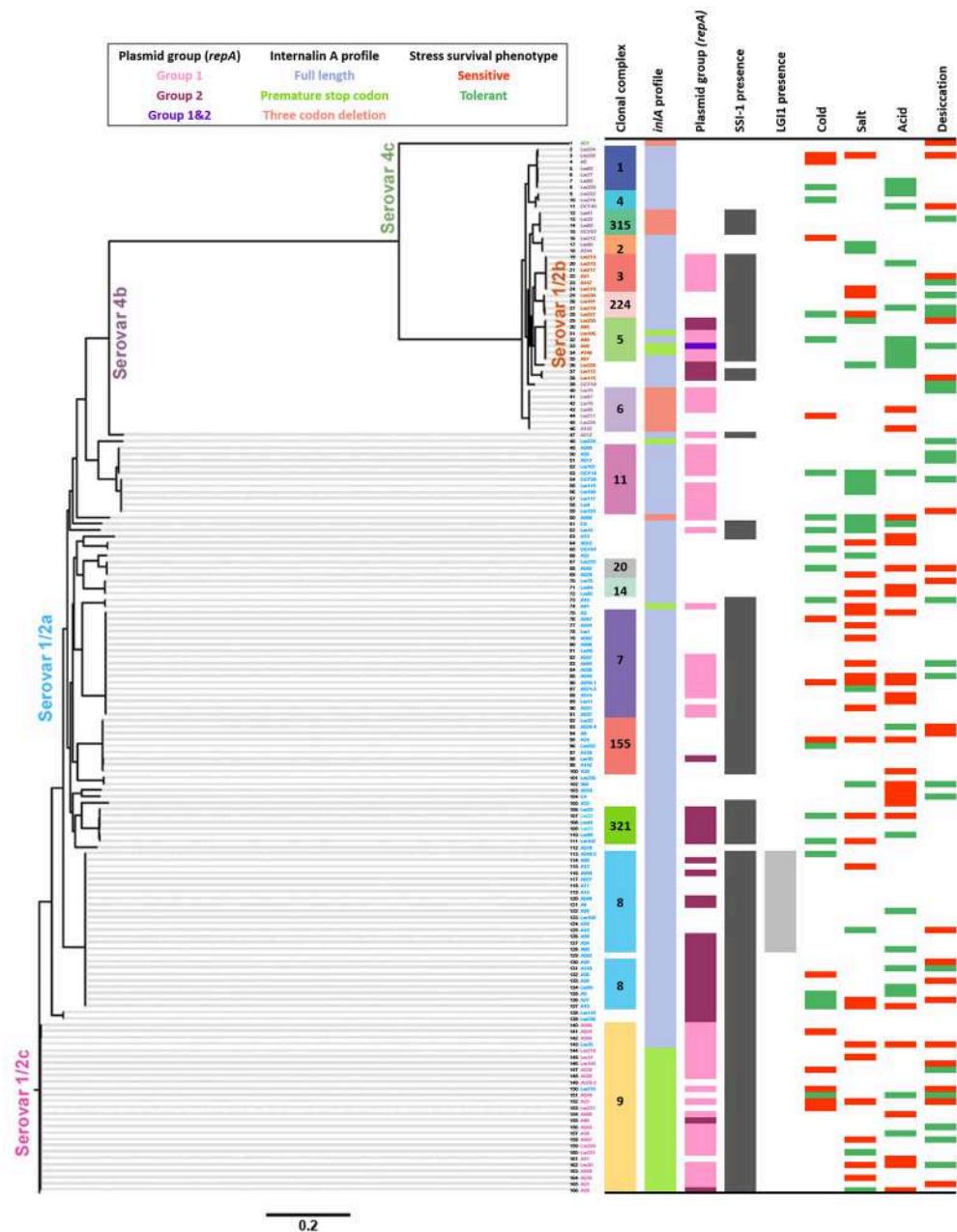


FIGURE 6 | Whole genome single nucleotide variant (SNV) phylogeny of 166 *L. monocytogenes* isolates and their associated genetic characteristics and stress tolerance phenotypes. The scale at the bottom indicates the substitutions per SNV.

However, they do not appear to be necessary for adequate cold growth as in the present study some 4b isolates were also classified as cold tolerant.

Salt Stress

LI strains have been shown to be more salt tolerant than LII strains (Bergholz et al., 2010) and serotype 4b strains to be more salt tolerant than serotype 1/2a and 1/2b strains (Van Der Veen et al., 2008; Bergholz et al., 2010; Ribeiro et al., 2014). In the

present study, no significant differences were found between the growth rates of different serotypes in 6% NaCl, however, five times as many 4b isolates were classified as salt tolerant as were classified as salt sensitive. Additionally, despite a 55% prevalence of serotype 1/2a isolates in our collection, isolates of this serotype accounted for 71% of salt tolerant isolates but also for 67% of salt sensitive isolates, indicating a broad range of salt tolerance among isolates from this serotype. These differences were found to be associated with specific 1/2a CCs, notably, CC7 isolates were

on average the most salt sensitive CC and significantly differed from the 1/2a CCs 8 and 11 with most 1/2a salt tolerant isolates belonging to CC11.

Acid Stress

LI isolates tolerated acid stress conditions significantly better than LII isolates. Specifically, serotype 1/2a showed higher sensitivity to acidity. This trend was also clear when the acid tolerance of individual CCs was investigated. Van Der Veen et al. (2008), also reported that 4b (LI) isolates had enhanced acid tolerance relative to 1/2a isolates. The authors hypothesized that the increased survival of 4b strains may in part be due to the presence of *ORF2110* which encodes a putative serine protease similar to HtrA. This protein has been shown to be important for growth at low pH, high osmolarity, and high temperatures (Wonderling et al., 2004; Stack et al., 2005; Wilson et al., 2006). Though this gene may contribute to the overall acid tolerance of 4b isolates, some acid sensitive 4b isolates were also identified. Similarly, despite serotype 1/2a isolates having relatively low acid tolerance overall, some 1/2a isolates were also acid tolerant, highlighting the importance of not overgeneralizing isolate phenotypes based on the trends seen for their sero- or sequence type.

Desiccation Stress

Desiccation tolerance reflects a bacteria's ability to survive on a surface for extended periods of time with little access to nutrients and water. As so, desiccation tolerance is believed to be associated with *L. monocytogenes*' ability to persist in food production plants (Vogel et al., 2010) and subsequently cross-contaminate food products. To date, surprisingly little research has been conducted regarding *L. monocytogenes*' desiccation survival and that which does exist has focused primarily on factors influencing the survival of a small number of isolates (Vogel et al., 2010; Hansen and Vogel, 2011; Takahashi et al., 2011; Hingston et al., 2013; Overney et al., 2016; Zoz et al., 2016). The current study is the first to our knowledge, to compare the desiccation tolerance of *L. monocytogenes* isolates from multiple serotypes. No significant differences were found between serotypes or CCs however, some prominent trends were observed. Serotypes 1/2c and 1/2b were on average the most desiccation tolerant, followed by 4b and 1/2a. More specifically, CC224 (1/2b) isolates had the highest levels of desiccation survival. Interestingly, a large listeriosis outbreak in Denmark, which resulted in 41 illnesses and 17 deaths, was linked to the consumption of deli meat contaminated with a CC224 strain (Kvistholm Jensen et al., 2016). Though there is not enough evidence to suggest that all CC224 strains are desiccation tolerant, it is possible that long-term desiccation survival may have contributed to the occurrence of this outbreak. Another interesting finding from the present study was that the most desiccation sensitive isolate, deviating more than 4.5 SD from the median, was a CC193 (serotype 1/2a) isolate. Since this was the only CC193 isolate in our collection, it would be interesting to analyze additional isolates from this CC to determine if this sequence type is associated with a high degree of desiccation sensitivity.

Certain Genetic Elements Are Associated with the Stress Tolerance of *L. monocytogenes*

Plasmids

The presumptive presence of a plasmid(s) was detected in 55% of our isolates which is comparable to other studies where rates of plasmid isolation have ranged from 0 to 79% with an overall average of around 30% (Perez-Diaz et al., 1982; Kolstad et al., 1992; Lebrun et al., 1992; Peterkin et al., 1992; McLauchlin et al., 1997). In agreement with earlier work, we also observed that plasmid DNA was more prevalent among LII isolates than LI isolates (Kolstad et al., 1992; Lebrun et al., 1992; McLauchlin et al., 1997; Margolles and de los Reyes-Gavilán, 1998; Orsi et al., 2011).

Kuenne et al. (2010) discovered that *L. monocytogenes* plasmids could be divided into two phylogenetic groups based on their *repA* sequences and that group 2 plasmids (77–83 kb) were larger than those belonging to group 1 (32–57 kb). Here plasmid *repA* sequences also formed two distinct phylogenetic groups and in agreement with Kuenne et al. (2010), group 2 plasmids were significantly larger (55–100 kb) than group 1 plasmids (26–88 kb). Notably, one serotype 1/2b isolate contained two plasmids of similar sizes (62 and 69 kb) but belonging to different *repA* groups. Though rare, the presence of two plasmids has been reported in other *Listeria* spp. isolates (Earnshaw and Lawrence, 1998; Margolles and de los Reyes-Gavilán, 1998).

Our results showed that among LII isolates, which exhibited higher rates of plasmid harborage than LI isolates, plasmid harborage was associated with significantly enhanced acid tolerance but also cold sensitivity. Studies have shown that plasmid harborage and subsequent replication increases the metabolic demands of cells, leading to decreased growth rate relative to plasmid-free strains (reviewed in Diaz Ricci and Hernández, 2000). However, depending on the genes contained on a plasmid, plasmid-harborage can also provide cells with a growth advantage when exposed to certain conditions. In this study, isolates with the larger *repA* group 2 plasmids were significantly more salt-tolerant than isolates that harbored the smaller *repA* group 1 plasmids, collectively suggesting that plasmid harborage may be a hindrance to *L. monocytogenes* during replication at low temperatures but provide an advantage when exposed to acid and salt stress conditions. Furthermore, the observation that isolates containing larger plasmids had higher levels of salt tolerance suggests that these plasmids may contain additional genes that are beneficial for adaptation to high osmolarity environments.

To date, plasmids acquired by *L. monocytogenes* have been shown to contain genes that confer resistance to benzalkonium chloride (*bcrABC*; Elhanafi et al., 2010; Rakic-Martinez et al., 2011; Katharios-Lanwermeyer et al., 2012), cadmium (*cadA2*, *cadAC*; Lebrun et al., 1992; Rakic-Martinez et al., 2011; Katharios-Lanwermeyer et al., 2012) and antibiotics including chloramphenicol, clindamycin, erythromycin, streptomycin, and tetracycline (Poyart-Salmeron et al., 1990; Hadorn et al., 1993). *Listeria* spp. plasmids also commonly contain several other uncharacterized efflux pumps (MDR, SMR, MATE; Gibson and

Parales, 2000; Masaoka et al., 2000; Boylan et al., 2006; Kuroda and Tsuchiya, 2009), as well as oxidative stress response genes (peroxidases, reductases; Kuenne et al., 2010) but their exact roles in stress tolerance have yet to be investigated. In other bacterial species, multidrug efflux pumps have been linked to stress response, virulence, and quorum sensing (reviewed in Li and Nikaido, 2009). Ma et al. (1995) reported that transcription of a MDR pump (*acrAB*) in *E. coli* increased in response to fatty acids, ethanol, high salt, and cellular transitioning into stationary phase. Among the putative plasmid associated contigs a wide variety of different genes were identified including those encoding cell surface proteins, lipoproteins, secretion pathways, heavy metal transporters, transcription regulators, general stress proteins (CplB, ClpL), NADH oxidoreductases, a glycine betaine transport permease (ProW), and the multidrug resistance proteins EbrA and EbrB among others. All group 2 plasmids shared a general secretion pathway protein and a cell surface protein. Other genes identified among many but not all group 2 plasmids included those which encoded DNA topoisomerase III, a membrane-bound protease, a NLP/P60 family lipoprotein, an NADH peroxidase, and a type IV secretory pathway. Further investigations are currently focusing on whether these genes or others with no known function contribute to *L. monocytogenes'* acid and salt tolerance.

Lastly, it should be highlighted that plasmids with 99–100% nucleotide identity were found in isolates from different serotypes, provinces (Alberta and British Columbia), and countries (Canada and Switzerland). One plasmid was observed in 26 isolates, which strongly suggests that *L. monocytogenes'* benefits from its presence. The occurrence of the same plasmid in multiple food-related isolates from different regions also suggests that bacteria are frequently transported between places of food production, possibly alongside imported raw materials. On the contrary, it is particularly interesting that other plasmids were conserved among specific CCs, serotypes, provinces, and countries.

Full Length *inlA*

Full length *inlA* profiles were observed among 92% of serotype 1/2a isolates, 83% of serotype 1/2b isolates, and 12% of serotype 1/2c isolates, reflecting what has been previously observed (Jonquieres et al., 1998; Jacquet et al., 2004; Rousseaux et al., 2004; Nightingale et al., 2005; Felicio et al., 2007; Orsi et al., 2007; Ragon et al., 2008). Additionally, 44% of 4b isolates contained a 3-codon deletion that unlike *inlA* PMSCs, is not associated with attenuated virulence (Kovacevic et al., 2013; Kanki et al., 2015).

The present study results showed that full length *inlA* profiles were more prevalent among cold, salt, and acid tolerant isolates compared to their sensitive counterparts. Also, isolates with full length *inlA* profiles were significantly more cold tolerant than isolates containing *inlA* PMSCs. Kovacevic et al. (2013) were the first to report that cold tolerant isolates more likely possess full length *inlA* than intermediate and cold sensitive isolates. This increased stress tolerance has now been shown to extend to salt and acid tolerance, making it reasonable to hypothesize that full length *inlA* may participate in *L. monocytogenes'* stress response. When bacteria are exposed to unfavorable conditions,

their cell envelope is the first line of defense. It is possible that the absence of cell wall anchored *InlA* proteins may alter cell-surface characteristics, leaving cells more susceptible to certain environmental stresses. Interestingly, only a small percentage of desiccation tolerant isolates contained full length *inlA* while serotype 4b isolates with full length *inlA* profiles had significantly impaired desiccation survival relative to those with a 3-codon deletion. Again, it is suspected that the structure of *inlA* may influence *L. monocytogenes'* desiccation tolerance, this time with the full length form possibly imparting a disadvantage. Other researchers have also detected associations between internalin mutations and certain phenotypes. In Hingston et al. (2015), *inlC* was identified as the interrupted gene in a desiccation tolerant transposon mutant and Franciosa et al. (2009) found that strains possessing a truncated *inlA* protein formed increased levels of biofilm. Similarly, transposon mutants containing an interrupted internalin A, B, or H gene, formed thicker biofilms relative to the wildtype (Piercey et al., 2016). Together, these findings along with those presented in this study, emphasize a need for more research regarding the potential roles of internalins in other processes other than virulence.

SSI-1

Stress survival islet 1 (SSI-1) is a five-gene cluster which has previously been shown via mutagenesis studies to enhance *L. monocytogenes* tolerance to acid, salt, and low temperature conditions (Cotter et al., 2005; Ryan et al., 2010). On the other hand, Arguedas-Villa et al. (2014) found no significant differences in cold tolerance between naturally occurring *L. monocytogenes* isolates with and without SSI-1. In the present study, it was found that SSI-1-positive isolates showed no enhanced cold, acid, salt, and desiccation stress tolerances relative to SSI-1 negative isolates. It is possible that any positive influence of SSI-1 on the stress tolerance of *L. monocytogenes'* may be masked by the presence of other genetic elements when comparing large collections of isolates as opposed to a mutant and its wildtype strain.

LGI1

LGI1 is a *Listeria* 50 kb genomic island that was first identified in Canadian CC8 isolates associated with a large 2008 listeriosis outbreak involving contaminated deli meats and resulting in 22 fatalities (Gilmour et al., 2010). Since then, LGI1 has been identified in other CC8 *L. monocytogenes* isolates from Canada (Kovacevic et al., 2013) but not from other countries (Althaus et al., 2014). In agreement with these studies, the presence of LGI1 was only detected in Canadian isolates from CC8. However, instead of all LGI1+ isolates being ST120 as previously reported, two novel CC8 STs (ST1022 and 1025) were also associated with LGI1 harborage. The conservation of LGI1 among Canadian isolates and its association with a fatal outbreak has led to heightened interest in the putative functions of the genes located on this island including those encoding putative type II and type IV secretion systems, pilus-like surface structures, a multidrug efflux pump homolog (EmrE), and an alternative sigma factor (Gilmour et al., 2010). Recently, Kovacevic et al. (2015) reported that deletion of LGI1 genes with putative efflux

(*emrE*), regulatory (*lmo1852*), and adhesion (*sel1*) functions, had no impact on the tolerance of *L. monocytogenes*' to acid, cold, or salt, but that deletion of *emrE* increased susceptibility to quaternary ammonium-based sanitizers. Based on these findings, it was investigated whether the presence or absence of the whole LGI1 island could be associated with stress tolerance differences between CC8 isolates however, no significant differences were identified. Consequently, LGI1 had no major influence on *L. monocytogenes*' ability to adapt to the food-related stresses evaluated in the present study. However, it is possible that the island contributes in other ways to the persistence of CC8 Canadian isolates in food processing environments in addition to the role of *emrE* in sanitizer resistance.

SNVs Associated with Stress Tolerance Phenotypes

An important finding from this study was that closely related isolates from within the same clonal complexes exhibited opposing stress tolerances, suggesting that minor genetic differences can also exert great impact on stress tolerance phenotypes. This was observed in a study by Hoffmann et al. (2013), where a single thymine deletion in the σ^A -like promoter region of *betL*, encoding an osmolyte transporter specific for betaine uptake, dramatically increased *betL* transcription, and hence the osmo- and chill-tolerance. Karatzas et al. (2003) reported that a spontaneous high hydrostatic pressure tolerant *L. monocytogenes* mutant of Scott A, contained a single codon deletion in *ctsR*, a negative regulator of several heat shock and general stress proteins, that also conferred increased thermo-tolerance and resistance to H₂O₂. Additionally, in Metselaar et al. (2015) a number of spontaneous acid tolerant mutants were found to contain SNVs in the ribosomal protein gene *rpsU*. None of these aforementioned mutations were detected in the present study.

Analyses to determine if unique SNVs could be detected among isolates from individual stress tolerance phenotype groups were also performed. Studies identifying the genetic basis of phenotypic traits using the variation within natural populations are known as a genome-wide association studies (GWAS). While GWAS have been effective for identifying mutations responsible for phenotypic traits in humans, the clonal nature of bacterial replication where mutations can reach a high frequency on a single genetic background, makes it difficult to distinguish mutations responsible for an observed phenotype (Read and Massey, 2014; Falush, 2016). As a result, bacterial molecular epidemiology has focused on identifying clonal lineages with particular phenotypic properties rather than identifying the specific genetic variants responsible. Recently, Earle et al. (2016) used a GWAS approach to successfully identify genes and genetic variants underlying resistance to 17 antimicrobials in over 3000 isolates of taxonomically diverse clonal and recombining bacteria. While these results show the potential of bacterial GWAS, antimicrobial resistance is usually gained during antimicrobial exposure and thus it is more likely that the traits evolve on multiple independent backgrounds making them easier to detect (Falush, 2016).

To date, the identification of mutations responsible for more complex phenotypes such as those evaluated in the present study, remain challenging.

Our analyses did not detect any unique SNVs among more than one isolate from the same stress tolerant group, suggesting homoplasy among stress tolerant phenotypes where mutations evolve independently to confer tolerance. On the contrary, a few different SNVs were identified among four or fewer isolates from the same stress sensitive groups. Notably, the cold sensitive phenotypes of two isolates may be associated with PMSCs detected in the σ^B regulator genes *rsbS* and *rsbV* (Voelker et al., 1995). In support of this hypothesis, deletion of *rsbV* in a *L. monocytogenes* mutagenesis study resulted in impaired cold stress tolerance (Chan et al., 2008). Three desiccation sensitive isolates also contained different PMSCs in *rsbS*, including one isolate which was also cold sensitive. An additional two desiccation sensitive isolates contained different PMSCs in another σ^B posttranscriptional regulator, *rsbU*. Given the importance of σ^B in *L. monocytogenes*' adaptation to several environmental stresses (Wiedmann et al., 1998; Kazmierczak et al., 2003), it is possible that these mutations contributed to the reduced desiccation tolerance of these isolates. In fact, in Huang et al. (2015) an *L. monocytogenes* *sigB* mutant demonstrated reduced desiccation survival relative to the wildtype strain.

Genomic Islands

In this study, specific genomic islands could not be exclusively associated with a particular stress tolerance phenotype. This is to be expected as genomic islands often contain virulence factors, and in general these are overrepresented in genomic islands as compared to the chromosome (Sui et al., 2009). In *L. monocytogenes* in particular, genomic islands have been associated with virulence, heavy metal resistance, and benzalkonium chloride efflux (Gilmour et al., 2010; Kuenne et al., 2010; Kovacevic et al., 2015). SSI-1 as described above, has been previously associated with *L. monocytogenes*' stress response, but was not significantly correlated with the stress tolerance phenotypes examined in the present study.

CONCLUSIONS

In summary, *L. monocytogenes*' tolerances to certain food-related stresses differs between serotypes as well as CCs with the latter being a better predictor of isolate salt and acid tolerance but not of cold and desiccation tolerance. To the best of our knowledge, this is the first study to evaluate the stress tolerance of different *L. monocytogenes* CCs. Other noteworthy findings include potential relationships between the presence of full length *inlA* and enhanced cold tolerance and the presence of a plasmid and enhanced acid tolerance. On the contrary, the presence of genomic islands including SSI-1 and LGI1, provided isolates with no noticeable advantages under the stresses evaluated in this study. Additional research is needed to confirm the potential roles of full length *inlA* and plasmid associated genes in *L. monocytogenes*' response to various stresses.

A whole genome SNV phylogeny of isolate assemblies identified a number of unique SNVs shared by up to four

stress sensitive isolates while no common SNVs were observed among stress tolerant isolates. More specifically, six isolates with sensitivity to cold and/or desiccation stress contained PMSCs in σ^B regulator genes (*rsbS*, *rsbU*, *rsbV*) that may be contributing to these phenotypes.

A number of novel genetic elements were also elucidated in this study including nine new *L. monocytogenes* STs, a new *inlA* PMSC, the absence of a cold stress associated gene (*lmo1078*) in 4b isolates, and several connections between *L. monocytogenes* CCs and the presence/absence or variations of specific genetic elements. For example, SSI-1 was detected in 100% of isolates from specific CCs, certain plasmid groups and sizes were conserved among isolates from the same CCs, and plasmids with 100% identity were found in isolates belonging to the same CCs but from very different geological areas. While our isolate collection represented of a number of *L. monocytogenes* CCs, some of which have been previously identified as common among food-related isolates, other CCs were less prevalent or absent from our study of Canadian and Swiss isolates. This highlights the regional prevalence of certain *L. monocytogenes* genotypes and emphasizes the need for more international collaborative studies.

Collectively, the results suggest that using whole genome sequencing to (1) determine the STs of *L. monocytogenes* food-related isolates and to (2) screen for the presence of genetic elements such as full length *inlA* and a plasmid(s), could help food processors and food agency investigators to quickly identify if isolates are likely to possess enhanced tolerances to certain stresses that may be facilitating their long-term survival/persistence in a food processing environment. The US FDA and CDC are rapidly making whole genome sequencing of foodborne bacterial pathogens a routine part of screening to help link illnesses to contaminated foods and to identify outbreaks earlier. While no one SNV was identified among isolates with

the same stress tolerant phenotype, increased sequencing of *L. monocytogenes* isolates in combination with stress tolerance profiling, will enhance the ability to identify genetic elements associated with more stress tolerant strains.

AUTHOR CONTRIBUTIONS

PH composed the paper and conducted all experiments and statistics. CL assembled and annotated the whole genome sequences. PH, JC, BD, and CB conducted the bioinformatics analyses. JC, VG, FB, TT, KA, LT, and SW provided guidance on study design, statistics, and assisted in writing the paper.

FUNDING

This work was funded by an investment agreement between Alberta Innovates—Bio Solutions and the University of British Columbia (FSC-12-030). PH was funded by an NSERC CGS D Scholarship. CB is supported by a fellowship from the Swiss National Science Foundation (P300-PA_164673) and a grant from the Société Académique Vaudoise.

ACKNOWLEDGMENTS

We thank Franco Pagotto and his team at the Listeriosis Reference Service at the Bureau of Microbial Hazards in Ottawa, Ontario for serotyping our isolates.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.00369/full#supplementary-material>

REFERENCES

- Althaus, D., Lehner, A., Brisse, S., Maury, M., Tasara, T., and Stephan, R. (2014). Characterization of *Listeria monocytogenes* strains isolated during 2011–2013 from human infections in Switzerland. *Foodborne Pathog. Dis.* 11, 753–758. doi: 10.1089/fpd.2014.1747
- Álvarez-Ordóñez, A., Fernández, A., López, M., Arenas, R., and Bernardo, A. (2008). Modifications in membrane fatty acid composition of *Salmonella Typhimurium* in response to growth conditions and their effect on heat resistance. *Int. J. Food Microbiol.* 123, 212–219. doi: 10.1016/j.ijfoodmicro.2008.01.015
- Annoos, B. A., Becker, L. A., Bayles, D. O., Labeda, D. P., and Wilkinson, B. J. (1997). Critical role of anteiso-C15:0 fatty acid in the growth of *Listeria monocytogenes* at low temperatures. *Appl. Environ. Microbiol.* 63, 3887–3894.
- Arguedas-Villa, C., Kovacevic, J., Allen, K. J., Stephan, R., and Tasara, T. (2014). Cold growth behaviour and genetic comparison of Canadian and Swiss *Listeria monocytogenes* strains associated with the food supply chain and human listeriosis cases. *Food Microbiol.* 40, 81–87. doi: 10.1016/j.fm.2014.01.001
- Arguedas-Villa, C., Stephan, R., and Tasara, T. (2010). Evaluation of cold growth and related gene transcription responses associated with *Listeria monocytogenes* strains of different origins. *Food Microbiol.* 27, 653–660. doi: 10.1016/j.fm.2010.02.009
- Aryani, D. C., Den Besten, H. M. W., Hazeleger, W. C., and Zwietering, M. H. (2015). Quantifying strain variability in modeling growth of *Listeria monocytogenes*. *Int. J. Food Microbiol.* 208, 19–29. doi: 10.1016/j.ijfoodmicro.2015.05.006
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comp. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021
- Baranyi, J., and Roberts, T. A. (1994). A dynamic approach to predicting bacterial growth in food. *Int. J. Food Microbiol.* 23, 277–294. doi: 10.1016/0168-1605(94)90157-0
- Barbosa, W. B., Cabedo, L., Wederquist, H. J., Sofos, J. N., and Schmidt, G. R. (1994). Growth variation among species and strains of *Listeria* in culture broth. *J. Food Prot.* 57, 765–769. doi: 10.4315/0362-028X-57.9.765
- Begot, C., Lebert, I., and Lebert, A. (1997). Variability of the response of 66 *Listeria monocytogenes* and *Listeria innocua* strains to different growth conditions. *Food Microbiol.* 14, 403–412. doi: 10.1006/fmic.1997.0097
- Bergholz, T. M., den Bakker, H. C., Fortes, E. D., Boor, K. J., and Wiedmann, M. (2010). Salt stress phenotypes in *Listeria monocytogenes* vary by genetic lineage and temperature. *Foodborne Pathog. Dis.* 7, 1537–1549. doi: 10.1089/fpd.2010.0624
- Borucki, M. K., and Call, D. R. (2003). *Listeria monocytogenes* serotype identification by PCR. *J. Clin. Microbiol.* 41, 5537–5540. doi: 10.1128/JCM.41.12.5537-5540.2003
- Boylan, J. A., Hummel, C. S., Benoit, S., Garcia-Lara, J., Treglown-Downey, J., Crane, E. J., et al. (2006). *Borrelia burgdorferi* bb0728 encodes a coenzyme A disulphide reductase whose function suggests a role in intracellular

- redox and the oxidative stress response. *Mol. Microbiol.* 59, 475–486. doi: 10.1111/j.1365-2958.2005.04963.x
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:10933404324
- Buchanan, R. L., Damert, W. G., Whiting, R. C., and van Schothorst, M. (1997). Use of epidemiologic and food survey data to estimate a purposefully conservative dose-response relationship for *Listeria monocytogenes* levels and incidence of listeriosis. *J. Food Prot.* 60, 918–922. doi: 10.4315/0362-028X-60.8.918
- Bulinski, A., Butkovsky, O., Shashkin, A., and Yaskov, P. (2011). Statistical methods of SNP data analysis with applications. arXiv:1106.4989
- Buncic, S., Avery, S. M., Rocourt, J., and Dimitrijevic, M. (2001). Can food-related environmental factors induce different behaviour in two key serovars, 4b and 1/2a, of *Listeria monocytogenes*? *Int. J. Food Microbiol.* 65, 201–212. doi: 10.1016/S0168-1605(00)00524-9
- Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., et al. (2005). Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.* 28, 171–182. doi: 10.1002/gepi.20041
- Cabedo, L., Picart i Barrot, L., and Teixidó i Canelles, A. (2008). Prevalence of *Listeria monocytogenes* and *Salmonella* in ready-to-eat food in Catalonia, Spain. *J. Food Prot.* 71, 855–859. doi: 10.4315/0362-028X-71.4.855
- Call, D. R., Borucki, M. K., and Besser, T. E. (2003). Mixed-genome microarrays reveal multiple serotype and lineage-specific differences among strains of *Listeria monocytogenes*. *J. Clin. Microbiol.* 41, 632–639. doi: 10.1128/JCM.41.2.632-639.2003
- CDC (2016a). *Listeria Outbreaks*. Available online at: <https://www.cdc.gov/listeria/outbreaks/>
- CDC (2016b). *The Listeria Whole Genome Sequencing Project*. Available online at: <https://www.cdc.gov/listeria/surveillance/whole-genome-sequencing.html>
- Chan, Y. C., and Wiedmann, M. (2008). Physiology and genetics of *Listeria monocytogenes* survival and growth at cold temperatures. *Crit. Rev. Food Sci. Nutr.* 49, 237–253. doi: 10.1080/10408390701856272
- Chan, Y. C., Hu, Y., Chaturongakul, S., Files, K. D., Bowen, B. M., Boor, K. J., et al. (2008). Contributions of two-component regulatory systems, alternative sigma factors, and negative regulators to *Listeria monocytogenes* cold adaptation and cold growth. *J. Food Prot.* 71, 420–425. doi: 10.4315/0362-028X-71.2.420
- Chassaing, D., and Auvray, F. (2007). The lmo1078 gene encoding a putative UDP-glucose pyrophosphorylase is involved in growth of *Listeria monocytogenes* at low temperature. *FEMS Microbiol. Lett.* 275, 31–37. doi: 10.1111/j.1574-6968.2007.00840.x
- Chavant, P., Martinie, B., Meylheuc, T., Bellon-Fontaine, M. N., and Hebraud, M. (2002). *Listeria monocytogenes* LO28: surface physicochemical properties and ability to form biofilms at different temperatures and growth phases. *Appl. Environ. Microbiol.* 68, 728–737. doi: 10.1128/AEM.68.2.728-737.2002
- Chen, Y., Ross, W. H., Scott, V. N., and Gombas, D. E. (2003). *Listeria monocytogenes*: low levels equal low risk. *J. Food Prot.* 66, 570–577. doi: 10.4315/0362-028X-66.4.570
- Chenal-Francisque, V., Lopez, J., Cantinelli, T., Caro, V., Tran, C., Leclercq, A., et al. (2011). Worldwide distribution of major clones of *Listeria monocytogenes*. *Emerg. Infect. Dis.* 17, 1110–1112. doi: 10.3201/eid1706.101778
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695
- Cotter, P. D., Ryan, S., Gahan, C. G., and Hill, C. (2005). Presence of GadD1 glutamate decarboxylase in selected *Listeria monocytogenes* strains is associated with an ability to grow at low pH. *Appl. Environ. Microbiol.* 71, 2832–2839. doi: 10.1128/AEM.71.6.2832-2839.2005
- Dalgaard, P., and Koutsoumanis, K. (2001). Comparison of maximum specific growth rates and lag times estimated from absorbance and viable count data by different mathematical models. *J. Microbiol. Methods* 43, 183–196. doi: 10.1016/S0167-7012(00)00219-0
- De Jesús, A. J., and Whiting, R. C. (2003). Thermal inactivation, growth, and survival studies of *Listeria monocytogenes* strains belonging to three distinct genotypic lineages. *J. Food Prot.* 66, 1611–1617. doi: 10.4315/0362-028X-66.9.1611
- De Lobel, L., Geurts, P., Baele, G., Castro-Giner, F., Kogevinas, M., and Van Steen, K. (2010). A screening methodology based on Random Forests to improve the detection of gene–gene interactions. *Eur. J. Hum. Genet.* 18, 1127–1132. doi: 10.1038/ejhg.2010.48
- Dhillon, B. K., Laird, M. R., Shay, J. A., Winsor, G. L., Lo, R., Nizam, F., et al. (2015). IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Res.* 43, W104–W108. doi: 10.1093/nar/gkv401
- Di Bonaventura, G., Piccolomini, R., Paludi, D., D'orio, V., Vergara, A., Conter, M., et al. (2008). Influence of temperature on biofilm formation by *Listeria monocytogenes* on various food-contact surfaces: relationship with motility and cell surface hydrophobicity. *J. Appl. Microbiol.* 104, 1552–1561. doi: 10.1111/j.1365-2672.2007.03688.x
- Diaz Ricci, J. C., and Hernández, M. E. (2000). Plasmid effects on *Escherichia coli* metabolism. *Crit. Rev. Biotechnol.* 20, 79–108. doi: 10.1080/0738855008984167
- Doumith, M., Buchrieser, C., Glaser, P., Jacquet, C., and Martin, P. (2004). Differentiation of the major *Listeria monocytogenes* serovars by multiplex PCR. *J. Clin. Microbiol.* 42, 3819–3822. doi: 10.1128/JCM.42.8.3819-3822.2004
- Dunn, K. A., Bielawski, J. P., Ward, T. J., Urquhart, C., and Gu, H. (2009). Reconciling ecological and genomic divergence among lineages of *Listeria* under an “extended mosaic genome concept”. *Mol. Biol. Evol.* 26, 2605–2615. doi: 10.1093/molbev/msp176
- Durack, J., Ross, T., and Bowman, J. P. (2013). Characterisation of the transcriptomes of genetically diverse *Listeria monocytogenes* exposed to hyperosmotic and low temperature conditions reveal global stress-adaptation mechanisms. *PLoS ONE* 8:e73603. doi: 10.1371/journal.pone.0073603
- Earle, S. G., Wu, C. H., Charlesworth, J., Stoesser, N., Gordon, N. C., Walker, T. M., et al. (2016). Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* 1:16041. doi: 10.1038/nmicrobiol.2016.41
- Earnshaw, A., and Lawrence, L. (1998). Sensitivity to commercial disinfectants, and the occurrence of plasmids within various *Listeria monocytogenes* genotypes isolated from poultry products and the poultry processing environment. *J. Appl. Microbiol.* 84, 642–648. doi: 10.1046/j.1365-2672.1998.00395.x
- Ebner, R., Stephan, R., Althaus, D., Brisse, S., Maury, M., and Tasara, T. (2015). Phenotypic and genotypic characteristics of *Listeria monocytogenes* strains isolated during 2011–2014 from different food matrices in Switzerland. *Food Control* 57, 321–326. doi: 10.1016/j.foodcont.2015.04.030
- Elhanafi, D., Dutta, V., and Kathariou, S. (2010). Genetic characterization of plasmid-associated benzalkonium chloride resistance determinants in a *Listeria monocytogenes* strain from the 1998–1999 outbreak. *Appl. Environ. Microbiol.* 76, 8231–8238. doi: 10.1128/AEM.02056-10
- Falush, D. (2016). Bacterial genomics: microbial GWAS coming of age. *Nat. Microbiol.* 1:16059. doi: 10.1038/nmicrobiol.2016.59
- Faul, F., Erdfelder, E., Buchner, A., and Lang, A. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149
- Faul, F., Erdfelder, E., Lang, A., and Buchner, A. (2007). G* Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146
- Felicio, M. T., Hogg, T., Gibbs, P., Teixeira, P., and Wiedmann, M. (2007). Recurrent and sporadic *Listeria monocytogenes* contamination in alheiras represents considerable diversity, including virulence-attenuated isolates. *Appl. Environ. Microbiol.* 73, 3887–3895. doi: 10.1128/AEM.02912-06
- Fenlon, D., Wilson, J., and Donachie, W. (1996). The incidence and level of *Listeria monocytogenes* contamination of food sources at primary production and initial processing. *J. Appl. Bacteriol.* 81, 641–650. doi: 10.1111/j.1365-2672.1996.tb03559.x
- Franciosa, G., Maugliani, A., Scalfaro, C., Floridi, F., and Aureli, P. (2009). Expression of internalin A and biofilm formation among *Listeria monocytogenes* clinical isolates. *Int. J. Immunopathol. Pharmacol.* 22, 183–193. doi: 10.1177/039463200902200121
- Gardner, S. N., Slezak, T., and Hall, B. G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* 31, 2877–2878. doi: 10.1093/bioinformatics/btv271
- Gibson, D. T., and Parales, R. E. (2000). Aromatic hydrocarbon dioxygenases in environmental biotechnology. *Curr. Opin. Biotechnol.* 11, 236–243. doi: 10.1016/S0958-1669(00)00090-2

- Gilmour, M. W., Graham, M., Van Domselaar, G., Tyler, S., Kent, H., Trout-Yakel, K. M., et al. (2010). High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. *BMC Genomics* 11:120. doi: 10.1186/1471-2164-11-120
- Hadorn, K., Hächler, H., Schaffner, A., and Kayser, F. (1993). Genetic characterization of plasmid-encoded multiple antibiotic resistance in a strain of *Listeria monocytogenes* causing endocarditis. *Eur. J. Hum. Genet.* 12, 928–937.
- Handa-Miya, S., Kimura, B., Takahashi, H., Sato, M., Ishikawa, T., Igarashi, K., et al. (2007). Nonsense-mutated inlA and prfA not widely distributed in *Listeria monocytogenes* isolates from ready-to-eat seafood products in Japan. *Int. J. Food Microbiol.* 117, 312–318. doi: 10.1016/j.ijfoodmicro.2007.05.003
- Hansen, L. T., and Vogel, B. F. (2011). Desiccation of adhering and biofilm *Listeria monocytogenes* on stainless steel: survival and transfer to salmon products. *Int. J. Food Microbiol.* 146, 88–93. doi: 10.1016/j.ijfoodmicro.2011.01.032
- Health Canada (2011). *Policy on Listeria monocytogenes in Ready-to-Eat Foods*. Available online at: http://www.hc-sc.gc.ca/fn-an/legislation/pol/policy-listeria-monocytogenes_2011-eng.php
- Hein, I., Klinger, S., Dooms, M., Flekna, G., Stessl, B., Leclercq, A., et al. (2011). Stress survival islet 1 (SSI-1) survey in *Listeria monocytogenes* reveals an insert common to *Listeria innocua* in sequence type 121 *L. monocytogenes* strains. *Appl. Environ. Microbiol.* 77, 2169–2173. doi: 10.1128/AEM.02159-10
- Hingston, P. A., Piercy, M. J., and Truelstrup Hansen, L. (2015). Genes associated with desiccation and osmotic stress in *Listeria monocytogenes* as revealed by insertional mutagenesis. *Appl. Environ. Microbiol.* 81, 5350–5362. doi: 10.1128/AEM.01134-15
- Hingston, P. A., Stea, E. C., Knöchel, S., and Hansen, T. (2013). Role of initial contamination levels, biofilm maturity and presence of salt and fat on desiccation survival of *Listeria monocytogenes* on stainless steel surfaces. *Food Microbiol.* 36, 46–56. doi: 10.1016/j.fm.2013.04.011
- Hoffmann, R. F., McLernon, S., Feeney, A., Hill, C., and Sleator, R. D. (2013). A single point mutation in the Listerial betL σA-dependent promoter leads to improved osmo- and chill-tolerance and a morphological shift at elevated osmolarity. *Bioengineered* 4, 401–407. doi: 10.4161/bioe.24094
- Hsiao, W. W., Ung, K., Aeschliman, D., Bryan, J., Finlay, B. B., and Brinkman, F. S. (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet.* 1:e62. doi: 10.1371/journal.pgen.0010062
- Huang, Y., Ells, T. C., and Hansen, L. T. (2015). Role of sigB and osmolytes in desiccation survival of *Listeria monocytogenes* in simulated food soils on the surface of food grade stainless steel. *Food Microbiol.* 46, 443–451. doi: 10.1016/j.fm.2014.09.007
- Jacquet, C., Doumith, M., Gordon, J. I., Martin, P. M., Cossart, P., and Lecuit, M. (2004). A molecular marker for evaluating the pathogenic potential of foodborne *Listeria monocytogenes*. *J. Infect. Dis.* 189, 2094–2100. doi: 10.1086/420853
- Jonquieres, R., Bierne, H., Mengaud, J., and Cossart, P. (1998). The inlA gene of *Listeria monocytogenes* LO28 harbors a nonsense mutation resulting in release of internalin. *Infect. Immun.* 66, 3420–3422.
- Junttila, J. R., Niemelä, S., and Hirn, J. (1988). Minimum growth temperatures of *Listeria monocytogenes* and non-haemolytic *Listeria*. *J. Appl. Bacteriol.* 65, 321–327. doi: 10.1111/j.1365-2672.1988.tb01898.x
- Kanki, M., Naruse, H., Taguchi, M., and Kumeda, Y. (2015). Characterization of specific alleles in InlA and PrfA of *Listeria monocytogenes* isolated from foods in Osaka, Japan and their ability to invade Caco-2 cells. *Int. J. Food Microbiol.* 211, 18–22. doi: 10.1016/j.ijfoodmicro.2015.06.023
- Karatzas, K. A., Wouters, J. A., Gahan, C. G., Hill, C., Abbe, T., and Bennik, M. H. (2003). The CtsR regulator of *Listeria monocytogenes* contains a variant glycine repeat region that affects piezotolerance, stress resistance, motility and virulence. *Mol. Microbiol.* 49, 1227–1238. doi: 10.1046/j.1365-2958.2003.03636.x
- Katharios-Lanwermeyer, S., Rakic-Martinez, M., Elhanafi, D., Ratani, S., Tiedje, J. M., and Kathariou, S. (2012). Coselection of cadmium and benzalkonium chloride resistance in conjugative transfers from nonpathogenic *Listeria* spp. to other *Listeriae*. *Appl. Environ. Microbiol.* 78, 7549–7556. doi: 10.1128/AEM.02245-12
- Kazmierczak, M. J., Mithoe, S. C., Boor, K. J., and Wiedmann, M. (2003). *Listeria monocytogenes* sigma B regulates stress response and virulence functions. *J. Bacteriol.* 185, 5722–5734. doi: 10.1128/JB.185.19.5722-5734.2003
- Klein, W., Weber, M. H., and Marahiel, M. A. (1999). Cold shock response of *Bacillus subtilis*: isoleucine-dependent switch in the fatty acid branching pattern for membrane adaptation to low temperatures. *J. Bacteriol.* 181, 5341–5349.
- Kolstad, J., Caugant, D. A., and Rørvik, L. M. (1992). Differentiation of *Listeria monocytogenes* isolates by using plasmid profiling and multilocus enzyme electrophoresis. *Int. J. Food Microbiol.* 16, 247–260. doi: 10.1016/0168-1605(92)90085-H
- Kovacevic, J., Arguedas-Villa, C., Wozniak, A., Tasara, T., and Allen, K. J. (2013). Examination of food chain-derived *Listeria monocytogenes* strains of different serotypes reveals considerable diversity in inlA genotypes, mutability, and adaptation to cold temperatures. *Appl. Environ. Microbiol.* 79, 1915–1922. doi: 10.1128/AEM.03341-12
- Kovacevic, J., Ziegler, J., Walecka-Zacharska, E., Reimer, A., Kitts, D. D., and Gilmour, M. W. (2015). Tolerance of *Listeria monocytogenes* to quaternary ammonium sanitizers is mediated by a novel efflux pump encoded by emR. *Appl. Environ. Microbiol.* 82, 939–953. doi: 10.1128/AEM.03741-15
- Kozak, J., Balmer, T., Byrne, R., and Fisher, K. (1996). Prevalence of *Listeria monocytogenes* in foods: incidence in dairy products. *Food Control* 7, 215–221. doi: 10.1016/S0956-7135(96)00042-4
- Kuenne, C., Voget, S., Pisichiarov, J., Oehm, S., Goesmann, A., Daniel, R., et al. (2010). Comparative analysis of plasmids in the genus *Listeria*. *PLoS ONE* 5:e12511. doi: 10.1371/journal.pone.0012511
- Kuroda, T., and Tsuchiya, T. (2009). Multidrug efflux transporters in the MATE family. *Biochim. Biophys. Acta* 1794, 763–768. doi: 10.1016/j.bbapap.2008.11.012
- Kvistholm Jensen, A., Nielsen, E. M., Bjorkman, J. T., Jensen, T., Muller, L., Persson, S., et al. (2016). Whole-genome sequencing used to investigate a nationwide outbreak of listeriosis caused by ready-to-eat delicatessen meat, Denmark, 2014. *Clin. Infect. Dis.* 63, 64–70. doi: 10.1093/cid/ciw192
- Lebrun, M., Loulergue, J., Chaslus-Dancla, E., and Audurier, A. (1992). Plasmids in *Listeria monocytogenes* in relation to cadmium resistance. *Appl. Environ. Microbiol.* 58, 3183–3186.
- Li, H. (2011). Improving SNP discovery by base alignment quality. *Bioinformatics* 27, 1157–1158. doi: 10.1093/bioinformatics/btr076
- Li, X., and Nikaido, H. (2009). Efflux-mediated drug resistance in bacteria. *Drugs* 69, 1555–1623. doi: 10.2165/11317030-000000000-00000
- Lianou, A., Stopforth, J. D., Yoon, Y., Wiedmann, M., and Sofos, J. N. (2006). Growth and stress resistance variation in culture broth among *Listeria monocytogenes* strains of various serotypes and origins. *J. Food Prot.* 69, 2640–2647. doi: 10.4315/0362-028X-69.11.2640
- Luber, P. (2011). The Codex Alimentarius guidelines on the application of general principles of food hygiene to the control of *Listeria monocytogenes* in ready-to-eat foods. *Food Control* 22, 1482–1483. doi: 10.1016/j.foodcont.2010.07.013
- Lunetta, K. L., Hayward, L., Segal, J., and Van Erdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.* 5:32. doi: 10.1186/1471-2156-5-32
- Ma, D., Cook, D. N., Alberti, M., Pon, N. G., Nikaido, H., and Hearst, J. E. (1995). Genes acrA and acrB encode a stress-induced efflux system of *Escherichia coli*. *Mol. Microbiol.* 16, 45–55. doi: 10.1111/j.1365-2958.1995.tb02390.x
- Margolles, A., and de los Reyes-Gavilán, C. G. (1998). Characterization of plasmids from *Listeria monocytogenes* and *Listeria innocua* strains isolated from short-ripened cheeses. *Int. J. Food Microbiol.* 39, 231–236. doi: 10.1016/S0168-1605(97)00132-3
- Martín, B., Perich, A., Gómez, D., Yangüela, J., Rodríguez, A., Garriga, M., et al. (2014). Diversity and distribution of *Listeria monocytogenes* in meat processing plants. *Food Microbiol.* 44, 119–127. doi: 10.1016/j.fm.2014.05.014
- Masaoka, Y., Ueno, Y., Morita, Y., Kuroda, T., Mizushima, T., and Tsuchiya, T. (2000). A two-component multidrug efflux pump, EbrAB, in *Bacillus subtilis*. *J. Bacteriol.* 182, 2307–2310. doi: 10.1128/JB.182.8.2307-2310.2000
- Maury, M. M., Tsai, Y. H., Charlier, C., Touchon, M., Chenal-Francisque, V., Leclercq, A., et al. (2016). Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nat. Genet.* 48, 308–313. doi: 10.1038/ng.3501
- McLauchlin, J., Hampton, M., Shah, S., Threlfall, E., Wieneke, A., and Curtis, G. (1997). Subtyping of *Listeria monocytogenes* on the basis of plasmid profiles and arsenic and cadmium susceptibility. *J. Appl. Microbiol.* 83, 381–388. doi: 10.1046/j.1365-2672.1997.00238.x
- Metselaar, K. I., den Besten, H. M., Boekhorst, J., van Hijum, S. A., Zwietering, M. H., and Abbe, T. (2015). Diversity of acid stress resistant variants of *Listeria*

- monocytogenes* and the potential role of ribosomal protein S21 encoded by *rpsU*. *Front. Microbiol.* 6:422. doi: 10.3389/fmicb.2015.00422
- Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Proc GCE*, 1–8. doi: 10.1109/gce.2010.5676129
- Moltz, A. G. (2005). Formation of biofilms by *Listeria monocytogenes* under various growth conditions. *J. Food Prot.* 68, 92–97. doi: 10.4315/0362-028X-68.1.92
- Mørøtø, T., and Langsrød, S. (2004). *Listeria monocytogenes*: biofilm formation and persistence in food-processing environments. *Biofilms* 1, 107–121. doi: 10.1017/S1479050504001322
- Nightingale, K. K., Ivy, R. A., Ho, A. J., Fortes, E. D., Njaa, B. L., Peters, R. M., et al. (2008). *inlA* premature stop codons are common among *Listeria monocytogenes* isolates from foods and yield virulence-attenuated strains that confer protection against fully virulent strains. *Appl. Environ. Microbiol.* 74, 6570–6583. doi: 10.1128/AEM.00997-08
- Nightingale, K. K., Windham, K., Martin, K. E., Yeung, M., and Wiedmann, M. (2005). Select *Listeria monocytogenes* subtypes commonly found in foods carry distinct nonsense mutations in *inlA*, leading to expression of truncated and secreted internalin A, and are associated with a reduced invasion phenotype for human intestinal epithelial cells. *Appl. Environ. Microbiol.* 71, 8764–8772. doi: 10.1128/AEM.71.12.8764-8772.2005
- Olier, M., Pierre, F., Lemaitre, J., Divies, C., Rousset, A., and Guzzo, J. (2002). Assessment of the pathogenic potential of two *Listeria monocytogenes* human faecal carriage isolates. *Microbiology* 148, 1855–1862. doi: 10.1099/00221287-148-6-1855
- Olier, M., Pierre, F., Rousseaux, S., Lemaitre, J. P., Rousset, A., Piveteau, P., et al. (2003). Expression of truncated Internalin A is involved in impaired internalization of some *Listeria monocytogenes* isolates carried asymptotically by humans. *Infect. Immun.* 71, 1217–1224. doi: 10.1128/IAI.71.3.1217-1224.2003
- Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *bioRxiv* 029827. doi: 10.1101/s13059-016-0997-x
- Orsi, R. H., Borowsky, M. L., Lauer, P., Young, S. K., Nusbaum, C., Galagan, J. E., et al. (2008). Short-term genome evolution of *Listeria monocytogenes* in a non-controlled environment. *BMC Genomics* 9:539. doi: 10.1186/1471-2164-9-539
- Orsi, R. H., den Bakker, H. C., and Wiedmann, M. (2011). *Listeria monocytogenes* lineages: genomics, evolution, ecology, and phenotypic characteristics. *Int. J. Med. Microbiol.* 301, 79–96. doi: 10.1016/j.ijmm.2010.05.002
- Orsi, R., Ripoll, D., Yeung, M., Nightingale, K., and Wiedmann, M. (2007). Recombination and positive selection contribute to evolution of *Listeria monocytogenes* *inlA*. *Microbiology* 153, 2666–2678. doi: 10.1099/mic.0.2007/007310-0
- Overney, A., Chassaing, D., Carpentier, B., Guillier, L., and Firmesse, O. (2016). Development of synthetic media mimicking food soils to study the behaviour of *Listeria monocytogenes* on stainless steel surfaces. *Int. J. Food Microbiol.* 238, 7–14. doi: 10.1016/j.ijfoodmicro.2016.08.034
- Parisi, A., Latorre, L., Normanno, G., Miccolupo, A., Fraccalvieri, R., Lorusso, V., et al. (2010). Amplified fragment length polymorphism and multi-locus sequence typing for high-resolution genotyping of *Listeria monocytogenes* from foods and the environment. *Food Microbiol.* 27, 101–108. doi: 10.1016/j.fm.2009.09.001
- Perez-Diaz, J., Vicente, M., and Baquero, F. (1982). Plasmids in *Listeria*. *Plasmid* 8, 112–118. doi: 10.1016/0147-619X(82)90049-X
- Peterkin, P. I., Gardiner, M., Malik, N., and Idziak, E. S. (1992). Plasmids in *Listeria monocytogenes* and other *Listeria* species. *Can. J. Microbiol.* 38, 161–164. doi: 10.1139/m92-027
- Piercey, M. J., Hingston, P. A., and Hansen, L. T. (2016). Genes involved in *Listeria monocytogenes* biofilm formation at a simulated food processing plant temperature of 15°C. *Int. J. Food Microbiol.* 223, 63–74. doi: 10.1016/j.ijfoodmicro.2016.02.009
- Poyart-Salmeron, C., Carlier, C., Trieu-Cuot, P., Courvalin, P., and Courtieu, A. (1990). Transferable plasmid-mediated antibiotic resistance in *Listeria monocytogenes*. *Lancet* 335, 1422–1426. doi: 10.1016/0140-6736(90)91447-I
- Ragon, M., Wirth, T., Hollandt, F., Lavenir, R., Lecuit, M., Le Monnier, A., et al. (2008). A new perspective on *Listeria monocytogenes* evolution. *PLoS Pathog.* 4:e1000146. doi: 10.1371/journal.ppat.1000146
- Rakic-Martinez, M., Drevets, D. A., Dutta, V., Katic, V., and Kathariou, S. (2011). *Listeria monocytogenes* strains selected on ciprofloxacin or the disinfectant benzalkonium chloride exhibit reduced susceptibility to ciprofloxacin, gentamicin, benzalkonium chloride, and other toxic compounds. *Appl. Environ. Microbiol.* 77, 8714–8721. doi: 10.1128/AEM.05941-11
- Read, T. D., and Massey, R. C. (2014). Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med.* 6:109. doi: 10.1186/s13073-014-0109-z
- Ribeiro, V., Mujahid, S., Orsi, R., Bergholz, T., Wiedmann, M., Boor, K., et al. (2014). Contributions of σ B and PrfA to *Listeria monocytogenes* salt stress under food relevant conditions. *Int. J. Food Microbiol.* 177, 98–108. doi: 10.1016/j.ijfoodmicro.2014.02.018
- Rissman, A. I., Mau, B., Biehl, B. S., Darling, A. E., Glasner, J. D., and Perna, N. T. (2009). Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* 25, 2071–2073. doi: 10.1093/bioinformatics/btp356
- Roldgaard, B. B., Andersen, J. B., Hansen, T. B., Christensen, B. B., and Licht, T. R. (2009). Comparison of three *Listeria monocytogenes* strains in a guinea-pig model simulating food-borne exposure. *FEMS Microbiol. Lett.* 291, 88–94. doi: 10.1111/j.1574-6968.2008.01439.x
- Rousseaux, S., Olier, M., Lemaitre, J. P., Piveteau, P., and Guzzo, J. (2004). Use of PCR-restriction fragment length polymorphism of *inlA* for rapid screening of *Listeria monocytogenes* strains deficient in the ability to invade Caco-2 cells. *Appl. Environ. Microbiol.* 70, 2180–2185. doi: 10.1128/AEM.70.4.2180-2185.2004
- Ryan, S., Begley, M., Hill, C., and Gahan, C. (2010). A five-gene stress survival islet (SSI-1) that contributes to the growth of *Listeria monocytogenes* in suboptimal conditions. *J. Appl. Microbiol.* 109, 984–995. doi: 10.1111/j.1365-2672.2010.04726.x
- Schwender, H., Zucknick, M., Ickstadt, K., Bolt, H. M., and GENICA network. (2004). A pilot study on the application of statistical classification procedures to molecular epidemiological data. *Toxicol. Lett.* 151, 291–299. doi: 10.1016/j.toxlet.2004.02.021
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153
- Shabala, L., Lee, S. H., Cannesson, P., and Ross, T. (2008). Acid and NaCl limits to growth of *Listeria monocytogenes* and influence of sequence of inimical acid and NaCl levels on inactivation kinetics. *J. Food Prot.* 71, 1169–1177. doi: 10.4315/0362-028X-71.6.1169
- Sorrells, K. M., Enigl, D. C., and Hatfield, J. R. (1989). Effect of pH, acidulant, time, and temperature on the growth and survival of *Listeria monocytogenes*. *J. Food Prot.* 52, 571–573. doi: 10.4315/0362-028X-52.8.571
- Stack, H. M., Sleator, R. D., Bowers, M., Hill, C., and Gahan, C. G. (2005). Role for HtrA in stress induction and virulence potential in *Listeria monocytogenes*. *Appl. Environ. Microbiol.* 71, 4241–4247. doi: 10.1128/AEM.71.8.4241-4247.2005
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033
- Sui, S. J. H., Fedynak, A., Hsiao, W. W., Langille, M. G., and Brinkman, F. S. (2009). The association of virulence factors with genomic islands. *PLoS ONE* 4:e8094. doi: 10.1371/journal.pone.0008094
- Takahashi, H., Kuramoto, S., Miya, S., and Kimura, B. (2011). Desiccation survival of *Listeria monocytogenes* and other potential foodborne pathogens on stainless steel surfaces is affected by different food soils. *Food Control* 22, 633–637. doi: 10.1016/j.foodcont.2010.09.003
- Team, R. C. (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: <http://www.R-project.org>
- Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15, 1–15. doi: 10.1186/s13059-014-0524-x
- US Food and Drug Administration (US FDA) (2016). *CPG Sec. 555.320 Listeria monocytogenes*. Available online at: <http://www.fda.gov/ICECI/ComplianceManuals/CompliancePolicyGuidanceManual/ucm136694.htm>
- Van Der Veen, S., Moelzaar, R., Abebe, T., and Wells-Bennik, M. H. (2008). The growth limits of a large number of *Listeria monocytogenes* strains at

- combinations of stresses show serotype-and niche-specific traits. *J. Appl. Microbiol.* 105, 1246–1258. doi: 10.1111/j.1365-2672.2008.03873.x
- Van Stelten, A., and Nightingale, K. K. (2008). Development and implementation of a multiplex single-nucleotide polymorphism genotyping assay for detection of virulence-attenuating mutations in the *Listeria monocytogenes* virulence-associated gene *inlA*. *Appl. Environ. Microbiol.* 74, 7365–7375. doi: 10.1128/AEM.01138-08
- Van Stelten, A., Simpson, J. M., Chen, Y., Scott, V. N., Whiting, R. C., Ross, W. H., et al. (2011). Significant shift in median guinea pig infectious dose shown by an outbreak-associated *Listeria monocytogenes* epidemic clone strain and a strain carrying a premature stop codon mutation in *inlA*. *Appl. Environ. Microbiol.* 77, 2479–2487. doi: 10.1128/AEM.02626-10
- Van Stelten, A., Simpson, J. M., Ward, T. J., and Nightingale, K. K. (2010). Revelation by single-nucleotide polymorphism genotyping that mutations leading to a premature stop codon in *inlA* are common among *Listeria monocytogenes* isolates from ready-to-eat foods but not human listeriosis cases. *Appl. Environ. Microbiol.* 76, 2783–2790. doi: 10.1128/AEM.02651-09
- Verheul, A., Russell, N. J., Van'T Hof, R., Rombouts, F. M., and Abee, T. (1997). Modifications of membrane phospholipid composition in nisin-resistant *Listeria monocytogenes* Scott A. *Appl. Environ. Microbiol.* 63, 3451–3457.
- Voelker, U., Voelker, A., Maul, B., Hecker, M., Dufour, A., and Haldenwang, W. G. (1995). Separate mechanisms activate sigma B of *Bacillus subtilis* in response to environmental and metabolic stresses. *J. Bacteriol.* 177, 3771–3780. doi: 10.1128/jb.177.13.3771-3780.1995
- Vogel, B. F., Hansen, L. T., Mordhorst, H., and Gram, L. (2010). The survival of *Listeria monocytogenes* during long term desiccation is facilitated by sodium chloride and organic material. *Int. J. Food Microbiol.* 140, 192–200. doi: 10.1016/j.ijfoodmicro.2010.03.035
- Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., et al. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinform.* 7:142. doi: 10.1186/1471-2105-7-142
- Walker, S., Archer, P., and Banks, J. G. (1990). Growth of *Listeria monocytogenes* at refrigeration temperatures. *J. Appl. Bacteriol.* 68, 157–162. doi: 10.1111/j.1365-2672.1990.tb02561.x
- Weber, M. H., Klein, W., Müller, L., Niess, U. M., and Marahiel, M. A. (2001). Role of the *Bacillus subtilis* fatty acid desaturase in membrane adaptation during cold shock. *Mol. Microbiol.* 39, 1321–1329. doi: 10.1111/j.1365-2958.2001.02322.x
- Wiedmann, M., Arvik, T. J., Hurley, R. J., and Boor, K. J. (1998). General stress transcription factor σB and its role in acid tolerance and virulence of *Listeria monocytogenes*. *J. Bacteriol.* 180, 3650–3656.
- Wilson, R. L., Brown, L. L., Kirkwood-Watts, D., Warren, T. K., Lund, S. A., King, D. S., et al. (2006). *Listeria monocytogenes* 10403S HtrA is necessary for resistance to cellular stress and virulence. *Infect. Immun.* 74, 765–768. doi: 10.1128/IAI.74.1.765-768.2006
- Wonderling, L. D., Wilkinson, B. J., and Bayles, D. O. (2004). The htrA (degP) gene of *Listeria monocytogenes* 10403S is essential for optimal growth under stress conditions. *Appl. Environ. Microbiol.* 70, 1935–1943. doi: 10.1128/AEM.70.4.1935-1943.2004
- Wu, S., Wu, Q., Zhang, J., Chen, M., and Guo, W. (2016). Analysis of multilocus sequence typing and virulence characterization of *Listeria monocytogenes* isolates from Chinese retail ready-to-eat food. *Front. Microbiol.* 7:168. doi: 10.3389/fmicb.2016.00168
- Yildiz, O., Aygen, B., Esel, D., Kayabas, U., Alp, E., Sumerkan, B., et al. (2007). Sepsis and meningitis due to *Listeria monocytogenes*. *Yonsei Med. J.* 48, 433–439. doi: 10.3349/ymj.2007.48.3.433
- Zhang, C., Zhang, M., Ju, J., Nietfeldt, J., Wise, J., Terry, P. M., et al. (2003). Genome diversification in phylogenetic lineages I and II of *Listeria monocytogenes*: identification of segments unique to lineage II populations. *J. Bacteriol.* 185, 5573–5584. doi: 10.1128/JB.185.18.5573-5584.2003
- Zoz, F., Iaconelli, C., Lang, E., Iddir, H., Guyot, S., Grandvalet, C., et al. (2016). Control of relative air humidity as a potential means to improve hygiene on surfaces: a preliminary approach with *Listeria monocytogenes*. *PLoS ONE* 11:e0148418. doi: 10.1371/journal.pone.0148418

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Hingston, Chen, Dhillon, Laing, Bertelli, Gannon, Tasara, Allen, Brinkman, Truelstrup Hansen and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Genome-Wide Association Study to Identify Diagnostic Markers for Human Pathogenic *Campylobacter jejuni* Strains

OPEN ACCESS

Edited by:

Sandra Torriani,
University of Verona, Italy

Reviewed by:

Heriberto Fernandez,
Austral University of Chile, Chile
Jinshui Zheng,
Huazhong Agricultural University,
China
Beatrix Stessl,
Veterinärmedizinische Universität
Wien, Austria

*Correspondence:

Eduardo N. Taboada
eduardo.taboada@canada.ca

†Present address:

Cody J. Buchanan,
Canadian Food Inspection Agency,
Canadian Science Centre for Human
and Animal Health, Winnipeg, MB,
Canada

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 27 January 2017

Accepted: 16 June 2017

Published: 30 June 2017

Citation:

Buchanan CJ, Webb AL,
Mutschall SK, Kruczakiewicz P,
Barker DOR, Hetman BM,
Gannon VPJ, Abbott DW,
Thomas JE, Inglis GD and
Taboada EN (2017) A Genome-Wide
Association Study to Identify
Diagnostic Markers for Human
Pathogenic *Campylobacter jejuni*
Strains. *Front. Microbiol.* 8:1224.
doi: 10.3389/fmicb.2017.01224

Cody J. Buchanan^{1,2†}, Andrew L. Webb¹, Steven K. Mutschall¹, Peter Kruczakiewicz¹, Dillon O. R. Barker^{1,2}, Benjamin M. Hetman¹, Victor P. J. Gannon¹, D. Wade Abbott³, James E. Thomas², G. Douglas Inglis³ and Eduardo N. Taboada^{1*}

¹ National Microbiology Laboratory at Lethbridge, Public Health Agency of Canada, Lethbridge, AB, Canada, ² Department of Biological Sciences, University of Lethbridge, Lethbridge, AB, Canada, ³ Lethbridge Research and Development Centre, Agriculture and Agri-Food Canada, Lethbridge, AB, Canada

Campylobacter jejuni is a leading human enteric pathogen worldwide and despite an improved understanding of its biology, ecology, and epidemiology, limited tools exist for identifying strains that are likely to cause disease. In the current study, we used subtyping data in a database representing over 24,000 isolates collected through various surveillance projects in Canada to identify 166 representative genomes from prevalent *C. jejuni* subtypes for whole genome sequencing. The sequence data was used in a genome-wide association study (GWAS) aimed at identifying accessory gene markers associated with clinically related *C. jejuni* subtypes. Prospective markers ($n = 28$) were then validated against a large number ($n = 3,902$) of clinically associated and non-clinically associated genomes from a variety of sources. A total of 25 genes, including six sets of genetically linked genes, were identified as robust putative diagnostic markers for clinically related *C. jejuni* subtypes. Although some of the genes identified in this study have been previously shown to play a role in important processes such as iron acquisition and vitamin B₅ biosynthesis, others have unknown function or are unique to the current study and warrant further investigation. As few as four of these markers could be used in combination to detect up to 90% of clinically associated isolates in the validation dataset, and such markers could form the basis for a screening assay to rapidly identify strains that pose an increased risk to public health. The results of the current study are consistent with the notion that specific groups of *C. jejuni* strains of interest are defined by the presence of specific accessory genes.

Keywords: *Campylobacter jejuni*, genome sequence, genome-wide association study, clinical association, molecular marker discovery, linkage analysis, molecular risk assessment

INTRODUCTION

Campylobacter jejuni is one of the leading causes of bacterial foodborne gastroenteritis in the world; it is estimated to be responsible for as much as 14% of all cases of diarrheal disease, translating to more than 400 million cases of campylobacteriosis annually (Duong and Konkel, 2009). In Canada, annual incidence rates nearing 30 cases per 100,000 individuals have been reported (Galanis, 2007),

although statistical models that account for unreported and undiagnosed cases suggest this rate could be as high as 447 cases per 100,000 individuals (Thomas et al., 2013). While a majority of cases are self-limiting, post-infection complications, such as Guillain-Barré syndrome can be life threatening (Nachamkin et al., 1998; Nachamkin, 2002). *Campylobacter jejuni* is commonly isolated from the gastrointestinal tract of many different wild and domesticated species, including companion animals and food animals such as poultry and cattle (Lastovica et al., 2014). Faecal contamination from carrier animals is considered to be a primary source of *C. jejuni* in the environment and on food products (Koenraad et al., 1997). This bacterium is highly prevalent in raw poultry meat and poultry by-products (Suzuki and Yamamoto, 2009; Williams and Oyarzabal, 2012), and the consumption and handling of contaminated poultry products is thought to be the primary source of exposure leading to human infection. Nonetheless, the epidemiology of campylobacteriosis is complex, with a large number of cases that appear to be sporadic (Silva et al., 2011), a range of animal and environmental reservoirs (Whiley et al., 2013), and multiple potential routes for the introduction of *C. jejuni* into the food chain as well as non-food-related pathways of exposure (Pintar et al., 2016).

Although epidemiological evidence suggests that not all *C. jejuni* strains or genetic lineages pose an equal risk to human health, our current understanding of *C. jejuni* subtype-dependent pathogenesis is incomplete. In contrast to other enteric pathogens, *C. jejuni* does not possess a number of the classical virulence factors (e.g., Type III or Type IV secretion systems, enterotoxins) found in other pathogens (Havelaar et al., 2009). Previous studies have identified genetic determinants that are important for *C. jejuni* pathogenicity (Dasti et al., 2010), but they are generally conserved across the species. Therefore, these factors have little predictive power for the identification of isolates with a higher propensity to cause disease in humans.

With the advent of inexpensive and high-throughput whole genome sequencing, Genome Wide Association Studies (GWAS) are increasingly being applied to bacterial genomics as tools for the identification of genetic markers associated with a phenotype or trait of interest (Read and Massey, 2014). GWAS represent a “top-down” approach to molecular marker discovery because the genomic content of “test” and “control” groups is compared and analyzed to identify genetic variation that is strongly associated with a given trait. This is in contrast to “bottom-up” approaches where individual genetic factors are manipulated to observe a phenotypic effect. The utility of GWAS lies in their ability to test many genetic factors in order to reveal associations with the phenotype of interest without *a priori* assumptions on the specific biological processes involved (Read and Massey, 2014). GWAS have been utilized to identify mutations and other polymorphisms associated with antibiotic resistance in *Mycobacterium tuberculosis* (Farhat et al., 2013), *Staphylococcus aureus* (Alam et al., 2014), and *Streptococcus pneumoniae* (Chewapreecha et al., 2014). In *Campylobacter*, GWAS have been used to identify genetic factors related to the Guillain-Barré Syndrome (Taboada et al., 2007), host adaptation in *C. jejuni* and *Campylobacter coli* (Sheppard et al., 2013), and

has recently been used to identify markers associated with the survival of *C. jejuni* in the poultry production chain (Yahara et al., 2016).

In this study, we have used isolates from the Canadian *Campylobacter* Comparative Genomic Fingerprinting Database (C3GFdb) to perform a GWAS aimed at identifying genetic determinants preferentially found among *C. jejuni* lineages associated with human disease. Comparative Genomic Fingerprinting (CGF) (Clark et al., 2012; Taboada et al., 2012) has been used as the primary tool for subtyping of *C. jejuni* isolates made available through a range of projects in Canada, including the FoodNet Canada sentinel surveillance program, the Canadian Integrated Program for Antimicrobial Surveillance, the Canadian Food Inspection Agency’s microbiological baseline survey of poultry, and several projects that incorporate human, food animal, wild animal, retail food, and environmental sampling activities. The C3GFdb currently contains subtyping data for 24,142 *Campylobacter* isolates from human ($n = 4,697$), animal ($n = 14,750$), and environmental ($n = 4,457$) sources from across Canada, representing 4,882 unique subtypes. It also contains basic epidemiological metadata for each isolate including host source, date and location, which facilitates contextualization of subtypes within the broader population structure of *C. jejuni* circulating in Canada.

The goal of the current study was to identify accessory genes with a statistically significant difference in carriage rates in two *C. jejuni* cohorts that differ in terms of their association with human campylobacteriosis. These genes could be used as diagnostic markers for molecular-based risk assessment and the rapid detection of *C. jejuni* isolates that pose the greatest risk to human health.

MATERIALS AND METHODS

Strain Selection

A total of 166 *C. jejuni* isolates representing 34 of the 100 most prevalent CGF subtypes circulating in Canada were selected from the C3GFdb for whole genome sequencing (Supplementary Table S1). The selected isolates and their respective subtypes represented approximately 31% (7,407/24,142) of all isolates in the database and over 55% (7,407/13,367) of the isolates from the 100 most prevalent CGF subtypes (Figure 1). They have been observed in multiple provinces, sources and hosts, and over multiple years, suggesting that they are endemic and in wide circulation. The dataset selected for WGS was comprised of 72 isolates from animals or retail meat, 54 isolates from environmental sources, and 40 isolates from human clinical cases (Table 1).

Genome Sequencing, Assembly, and Annotation

Sequencing was conducted at Canada’s Michael Smith Genome Sciences Centre, BC, Canada using the Illumina HiSeq 2000 platform. Whole genome sequence data for this study has been deposited in the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) under

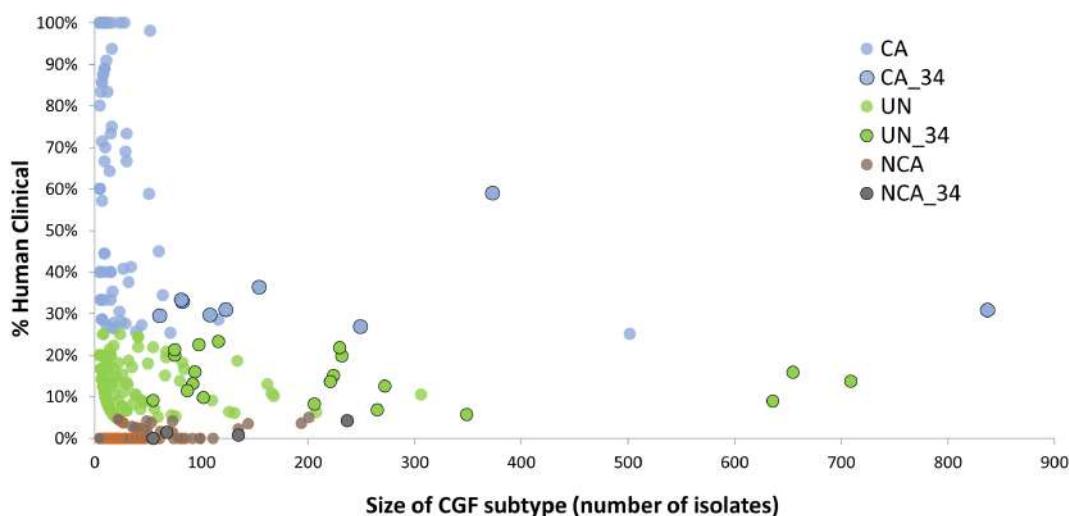


FIGURE 1 | Identification of CGF subtypes for GWAS analysis of Clinically-Associated (CA) vs. Non-Clinically-Associated (NCA) *Campylobacter jejuni* subtypes. The C3GFdb was used to identify 166 *C. jejuni* isolates for whole genome sequencing from 34 highly prevalent CGF subtypes (black outline) that together account for nearly 31% of all isolates in the database, and over 55% of all isolates from the 100 most prevalent CGF subtypes circulating in Canada. These subtypes exhibit differences in their association with human campylobacteriosis, and sequence data from representative isolates was used in a genome-wide association analysis aimed at identifying accessory genes associated with clinically relevant *C. jejuni* subtypes.

the BioProject PRJNA368735. Draft *de novo* genome assembly of paired-end reads was performed using SPADES v.2.4.0 (Bankevich et al., 2012) with pre-assembly BayesHammer read correction, default k-mer size testing options, and post-assembly Burrows Wheeler Aligner mismatch correction. Contigs with low coverage or shorter than 500 bp were removed from all subsequent analyses. Genome assembly quality was assessed using QUAST v.2.1 (Gurevich et al., 2013). Prediction of Open Reading Frames (ORFs) and annotation was performed using the PROKKA pipeline v.1.5.2 (Seemann, 2014) using a custom database of non-redundant gene sequences representing five complete and well-annotated *C. jejuni* reference genomes available from NCBI (Supplementary Table S2).

Definition of a *C. jejuni* Reference Pan-Genome for the Dataset

Predicted ORFs were queried using a reciprocal best hit approach (Moreno-Hagelsieb and Latimer, 2008; Ward and Moreno-Hagelsieb, 2014) with BLAST v 2.2.29 (Camacho et al., 2009) in order to define a reference pan-genome, the non-redundant set of genes for a set of genome sequences (Méric et al., 2014). Paired BLAST queries were treated as *orthologous* if they shared $\geq 80\%$ sequence identity and $\geq 50\%$ alignment coverage and a single exemplar was included in the pan-genome. The pan-genome defined using this process was used in the subsequent GWAS.

Genome Wide Association Study

Carriage across the dataset of all genes representing the pan-genome was assessed by BLAST analysis. The nucleotide sequence of each gene was queried against the 166 draft genome assemblies using Blastn. Genes were considered to be *present* if a hit representing $\geq 80\%$ sequence identity over $\geq 50\%$ of the length

of the query gene was found and considered *absent* otherwise. In order to facilitate statistical comparison, subtypes were defined as either Non-Clinically Associated (NCA; $\leq 5\%$ human clinical isolates), Undefined (UN; 5–25% human clinical isolates), or Clinically Associated (CA; $\geq 25\%$ human clinical isolates). The statistical significance of each gene ($p < 0.05$) was defined based on its carriage rate in the CA and NCA cohorts and was computed using Fisher's Exact test statistic in GenomeFisher¹; p -values were adjusted for multiple testing using the method of Holm (Holm, 1979; Aickin and Gensler, 1996). Statistically significant genes were subjected to further analysis and validation as outlined below.

In Silico Validation of Putative Diagnostic Marker Genes

In order to select markers with the highest potential for downstream assay development, candidate genes identified by the GWAS analysis were filtered in a stepwise process according to the following conditions: (1) complete absence in the NCA cohort and presence in $\geq 50\%$ of CA genomes; (2) high sequence identity ($> 99\%$) and complete, or near complete, conservation of sequence length ($> 90\%$) in the corresponding orthologous gene, when present, among a set of reference genomes (Table 2); and (3) statistical significance ($p < 0.05$) when the NCA cohort was compared to a combined CA+UN cohort, in which the UN (i.e., undefined) genomes were treated as CA and pooled with the CA genomes. Genes that passed all criteria were selected for *in silico* validation using a larger set of genome sequences. This validation dataset was created by combining genomes sequenced in house as part of current or previous studies ($n = 325$) and additional genomes acquired from public repositories

¹<https://bitbucket.org/peterk87/genomefisher/wiki/Home>

($n = 3,955$). Publicly available genomes were restricted to those with available epidemiological data (e.g., sample source, country of origin, etc.). To facilitate assignment into NCA, UN, and CA cohorts, *in silico* CGF was performed on these genomes using MIST (Kruczkiewicz et al., 2013), with a concordance between CGF profiles predicted *in silico* and those generated in the laboratory estimated to be 96.8% on a subset of 325 isolates (12,583/13,000 matching loci; data not shown); only genomes from CGF subtypes previously observed in the C3GFdb were retained in the validation set ($n = 3,902$). Each genome was designated to its respective cohort based on the corresponding epidemiological data of the *in silico* CGF subtype. Finally, the putative diagnostic genes identified by the GWAS using the

original set of 166 genomes were tested for statistical significance with the expanded cohorts. The combinatorial effect of different subsets of markers was also assessed to determine if a reduced number of markers could be applied to detect clinically related *C. jejuni* subtypes without a subsequent loss of sensitivity.

RESULTS AND DISCUSSION

Genome Sequencing, Assembly, and Annotation

The quality of the *de novo* assembly of the 166 genomes selected as representatives of 34 highly prevalent CGF subtypes

TABLE 1 | Epidemiological characteristics of 34 CGF subtypes targeted for whole-genome sequencing based on the Canadian Campylobacter Comparative Genomic Fingerprinting Database (C3GFdb).

CGF Subtype	Cohort ¹	Cluster Size ²	Cluster Rank ³	Proportion of isolates in subtype (%) ⁴			
				H	A	E	U
0169.001.002	CA	837	1	30.8%	62.7%	6.3%	0.1%
0695.006.001	UN	709	2	13.7%	80.4%	5.9%	0.0%
0083.001.002	UN	655	3	15.9%	83.2%	0.8%	0.2%
0926.002.001	UN	636	4	9.0%	74.2%	16.8%	0.0%
0044.003.001	CA	373	6	59.0%	40.5%	0.5%	0.0%
0957.001.001	UN	349	7	5.7%	69.6%	24.6%	0.0%
0853.011.001	UN	272	9	12.5%	87.1%	0.4%	0.0%
0882.005.001	UN	265	10	6.8%	81.1%	9.4%	2.6%
0982.001.002	CA	249	11	26.9%	68.3%	4.8%	0.0%
0811.009.002	NCA	237	12	4.2%	43.9%	51.9%	0.0%
0735.005.001	UN	232	13	19.8%	66.8%	13.4%	0.0%
0253.004.001	UN	230	14	21.7%	75.2%	3.0%	0.0%
0960.007.001	UN	224	15	15.2%	76.8%	5.4%	2.7%
0731.001.005	UN	221	16	13.6%	81.9%	4.5%	0.0%
0923.002.001	UN	206	18	8.3%	61.7%	30.1%	0.0%
0269.004.001	CA	154	24	36.4%	63.6%	0.0%	0.0%
0811.008.001	NCA	135	26.5	0.7%	45.9%	53.3%	0.0%
0173.004.001	CA	123	31	30.9%	57.7%	11.4%	0.0%
0173.002.004	UN	116	32.5	23.3%	76.7%	0.0%	0.0%
0933.004.002	CA	108	36	29.6%	65.7%	4.6%	0.0%
0893.001.001	UN	102	37	9.8%	82.4%	7.8%	0.0%
0933.008.001	UN	98	40	22.4%	75.5%	2.0%	0.0%
0949.001.002	UN	94	41	16.0%	72.3%	11.7%	0.0%
0960.003.002	UN	92	42.5	13.0%	67.4%	19.6%	0.0%
0904.002.002	UN	87	44	11.5%	74.7%	12.6%	1.1%
0103.001.002	CA	82	48.5	32.9%	67.1%	0.0%	0.0%
0077.001.003	CA	81	51	33.3%	66.7%	0.0%	0.0%
0238.007.002	UN	75	55.5	20.0%	80.0%	0.0%	0.0%
0260.007.001	UN	75	55.5	21.3%	78.7%	0.0%	0.0%
0844.001.001	NCA	68	63	1.5%	23.5%	75.0%	0.0%
0253.001.002	CA	61	69.5	29.5%	68.9%	0.0%	1.6%
0535.001.003	UN	55	76.5	9.1%	36.4%	54.5%	0.0%
0817.003.001	NCA	55	76.5	0.0%	20.0%	80.0%	0.0%
0083.007.001	CA	51	81	58.8%	41.2%	0.0%	0.0%

¹Cohorts: Clinically Associated (CA); Non-Clinically Associated (NCA); Undefined (UN). ²Number of isolates observed with the CGF subtype in the C3GFdb. ³Rank of CGF subtype (based cluster size) in the C3GFdb. ⁴Proportion of isolates in the subtype from Human (H), Animal (A), Environmental (E), and Unknown (U) sources.

in Canada was assessed using QUAST (Gurevich et al., 2013). The average number of reads produced for each genome was 4,161,271 ($\pm 1,223,304$), for an average coverage of 253 \times ($\pm 74.7\times$). Individual genome assemblies had an average of 67 (± 27) contigs and an N75 of 34,631 bp ($\pm 13,815$ bp). All genome assemblies had additional parameters in range with what has typically been observed for *C. jejuni*. The average assembly length (1,660,986 \pm 51,283.5 bp), predicted ORFs (1,719 \pm 71), and %G+C (30.4 \pm 0.13%) were typical of *C. jejuni* genome assemblies available in the public domain. Annotation of the

166 draft genomes from this study using the PROKKA pipeline (Seemann, 2014) resulted in the identification of 291,502 ORFs. The genome of strain NCTC 11168, which has been completely sequenced (Parkhill et al., 2000), was included in the analysis as a control to assess the completeness of the ORF prediction and annotation process. The original annotation of NCTC 11168 predicted 1,654 ORFs, while a subsequent re-annotation predicted 1,643 ORFs (Gundogdu et al., 2007); in our analysis, the PROKKA pipeline predicted 1,659 ORFs. This small discrepancy is related to the advanced curation used in the re-annotation

TABLE 2 | Significant genes observed after GWAS analysis of genome sequences from representative Clinically Associated (CA) and Non-Clinically Associated (NCA) *C. jejuni* subtypes.

Marker	<i>p</i> -value ¹		11168 Ortholog	Gene name	Function	Linkage group
	Raw	Holm-corrected ²				
11168_00051	4.29E-10	8.39E-07	<i>Cj0055c</i>		Hypothetical protein	LG1
11168_00052	5.28E-10	1.03E-06	<i>Cj0056c</i>		Hypothetical protein	
11168_00169	3.36E-11	6.61E-08	<i>Cj0177</i>		Putative iron transport protein	LG2
11168_00170	3.36E-11	6.61E-08	<i>Cj0178</i>		Putative TonB-dependent outer membrane receptor	
11168_00171	3.36E-11	6.60E-08	<i>Cj0179</i>	<i>exbB1</i>	Biopolymer transport protein	
11168_00172	3.36E-11	6.60E-08	<i>Cj0180</i>	<i>exbD1</i>	Biopolymer transport protein	
11168_00173	3.36E-11	6.60E-08	<i>Cj0181</i>	<i>tonB1</i>	TonB transport protein	
11168_00230	6.12E-19	1.21E-15	<i>Cj0246c</i>		Putative MCP-domain signal transduction protein	
11168_00243	6.48E-34	1.28E-30	<i>Cj0259</i>	<i>pyrC</i>	Putative dihydroorotate	LG3
11168_00244	3.10E-27	6.14E-24	<i>Cj0260c</i>		Small hydrophobic protein	
11168_00248	6.57E-25	1.30E-21	<i>Cj0264c</i>		Putative molybdopterin containing oxidoreductase	LG4
11168_00249	6.57E-25	1.30E-21	<i>Cj0265c</i>		Putative cytochrome C-type haem-binding Periplasmic protein	
11168_00277	1.30E-17	2.57E-14	<i>Cj0295</i>		Putative acetyltransferease	LG5
11168_00278	1.35E-18	2.66E-15	<i>Cj0296c</i>	<i>panD</i>	Aspartate 1-decarboxylase precursor	
11168_00279	1.35E-18	2.66E-15	<i>Cj0297c</i>	<i>panC</i>	Pantoate-beta-alanine ligase	
11168_00280	1.35E-18	2.66E-15	<i>Cj0298c</i>	<i>panB</i>	3-methyl-2-oxobutanoate hydroxymethyltransferase	
11168_00281	1.09E-16	2.15E-13	<i>Cj0299</i>		Putative periplasmic beta-lactamase	
11168_00703	6.98E-24	1.38E-20	<i>Cj0731</i>		Putative ABC transport system permease	
11168_00718	3.36E-11	6.59E-08	<i>Cj0753c</i>	<i>tonB3</i>	TonB transport protein	LG6
11168_00719	3.36E-11	6.59E-08	<i>Cj0755</i>	<i>cfrA</i>	Ferric enterobactin uptake receptor	
11168_01072	4.90E-11	9.59E-08	<i>Cj1122c</i>		Putative integral membrane protein.	
11168_01201	6.12E-19	1.21E-15	<i>Cj1255</i>		Putative isomerase	
11168_01309	5.30E-15	1.04E-11	<i>Cj1365c</i>		Putative secreted serine protease	
11168_01519	4.29E-10	8.39E-07	<i>Cj1585c</i>		Putative oxidoreductase	
11168_01610	4.29E-10	8.38E-07	<i>Cj1679</i>		Hypothetical protein	
06_2866_00597	6.89E-28	1.36E-24			Di-/tripeptide transporter	
06_7515_00723	4.19E-16	8.24E-13			Prophage Lp2 protein 6	
07_0675_00227	2.62E-11	5.15E-08		<i>tetO</i>	Elongation factor G	

¹*p*-value based on 2-tailed Fisher's Exact Test. ²*p*-values were adjusted using the Holm-correction (Holm, 1979).

of NCTC 11168, which resulted in the merging and removal of coding sequences belonging to pseudogenes and phase variable genes. The pan-genome established using this dataset consisted of 3,358 unique ORFs, of which 1,377 were present in all genomes (i.e., core genes) and 1,981 were present in a varying number of genomes (i.e., accessory genes).

Genome Wide Association Study

Of the 166 *C. jejuni* isolates selected for this study, 35 were assigned to the NCA cohort and represented four different CGF subtypes, 80 were assigned to the UN cohort and represented 20 CGF subtypes, and 51 were assigned to the CA cohort and represented ten CGF subtypes (Table 1). A GWAS was performed in order to identify accessory genes with a biased distribution in CA and NCA cohorts. Although in principle GWAS can be used to identify genetic variation ranging from SNPs to indels involving multiple genes, we chose to focus on accessory genes, as they have excellent potential for the development of rapid, robust, and inexpensive PCR-based diagnostic assays for screening of large numbers of strains. At the same time, it is important to note that other forms of genetic variation may represent valuable targets for tracking strains of interest. Recently, Clark et al. (Clark et al., 2016) showed that large-scale chromosomal inversion could be used to distinguish a subset of outbreak-associated isolates from epidemiologically unrelated co-circulating isolates.

In total, 595 genes showed statistically significant differences in carriage between NCA and CA cohorts ($p \leq 0.05$) (Figure 2). Of these, 71 genes were completely absent from the NCA cohort but were present in at least $\geq 50\%$ of isolates in the CA cohort (Condition 1), and 63 of these genes also maintained high sequence identity (>99%) and near complete sequence coverage (>90%) compared to their respective reference genes (Condition 2). Of these, 28 continued to exhibit robust statistical significance when the NCA cohort was compared to a pooled cohort comprised of all UN and CA genomes (Condition 3). These include six sets of genes that appear to be found in linkage groups (Table 2), with members of each linkage group possessing similar rates of carriage in the dataset. Since linked genes, which are located adjacently on the chromosome, tend to be functionally related and are typically transmitted as a functional unit (Muley and Acharya, 2013), it is likely that their identification in this study was not due to spurious statistical signal.

Among the linkage groups observed in the GWAS were two sets of genes responsible for encoding iron acquisition systems. We observed that genes encoding both the *TonB1*-mediated *Cj0178* (LG2; *Cj0177-Cj0181*) and the *TonB3*-mediated *CfrA* (LG6; *Cj0753c/Cj0755*) iron acquisition systems were significantly associated with *C. jejuni* isolates from clinically related CGF subtypes. As is the case in most pathogens, iron acquisition is considered to be a virulence determinant in *C. jejuni* and has been linked to successful colonization *in vivo* (Kim et al., 2003; Palyada et al., 2004; Naikare et al., 2006). *CfrA* has been shown to be capable of transporting a wide variety of structurally different siderophores, which may contribute to the ability of isolates with these genes to colonize a wide variety of hosts/niches (Naikare et al., 2013).

Another linkage group associated with CA and UN subtypes was comprised of genes that encode the pantothenate (vitamin B₅) biosynthesis pathway and β -lactam antibiotic resistance. LG5 encompasses a total of five genes, including a putative acetyltransferase (*Cj0295*), the *panBCD* operon (*Cj0296c-Cj0298c*), which encodes for the pantothenate (vitamin B₅) biosynthesis pathway, as well as the gene *bla_{OXA-61}* (*Cj0299*), which encodes a protein that confers resistance to β -lactam antibiotics. These genes were recently implicated in host adaptation in *C. jejuni* and *C. coli*, where they were found to be more strongly associated with cattle-specific lineages relative to chicken-specific lineages, possibly as a result of selective pressures created by contemporary and geographically dependent agricultural practices (Sheppard et al., 2013). Although it is generally recognized that chickens are a primary source of human exposure leading to infection, we observed strong statistical signal among CA subtypes for genes previously identified as cattle-associated (Sheppard et al., 2013). Sheppard et al. (2013) suggested that maintenance of these genes in chickens, albeit at a reduced rate, may facilitate rapid-host switching as part of a host-generalist strategy. Moreover, we have observed that a majority of the most prevalent clinically related CGF subtypes, many of which are represented in our GWAS dataset, are associated with both cattle and chickens. This is consistent with the possible role of cattle as an important reservoir for strains that go on to contaminate the chicken production system, ultimately leading to human cases of campylobacteriosis. As this manuscript was being readied for publication, GWAS was used to identify several loci that could be used as “host-segregating” epidemiological markers for source attribution (Thépault et al., 2017). Interestingly, one of the loci (*Cj0260c*) was also identified in our analysis. Thus, while our data suggests that presence of this gene is strongly associated with human clinical isolates, data from the study by Thépault et al. further suggests the allelic information appears highly predictive of host source.

In Silico Validation of Putative Diagnostic Marker Genes

Population structure has been identified as a potential confounding factor in GWAS analyses, in that statistically significant associations may ultimately be due to oversampling of certain subpopulations rather than with the phenotypic trait under investigation (Read and Massey, 2014). Since the focus of the current study was the examination of prevalent *C. jejuni* subtypes in Canada in the context of population structure, it was necessary to exclude the possibility that the markers we identified represent a biased distribution resulting from oversampling within certain lineages in the population. The large-scale marker validation that we performed using available WGS data included a dataset comprised of genomes largely from the United Kingdom (3,871/4,280; 90%) and Canada (327/4,280; 8%), and an overwhelming majority of isolates were recovered from human clinical sources (3,559/4,280; 83%), while those from animal (626/4,280; 15%) and environmental (95/4,280; 2%) sources comprised the

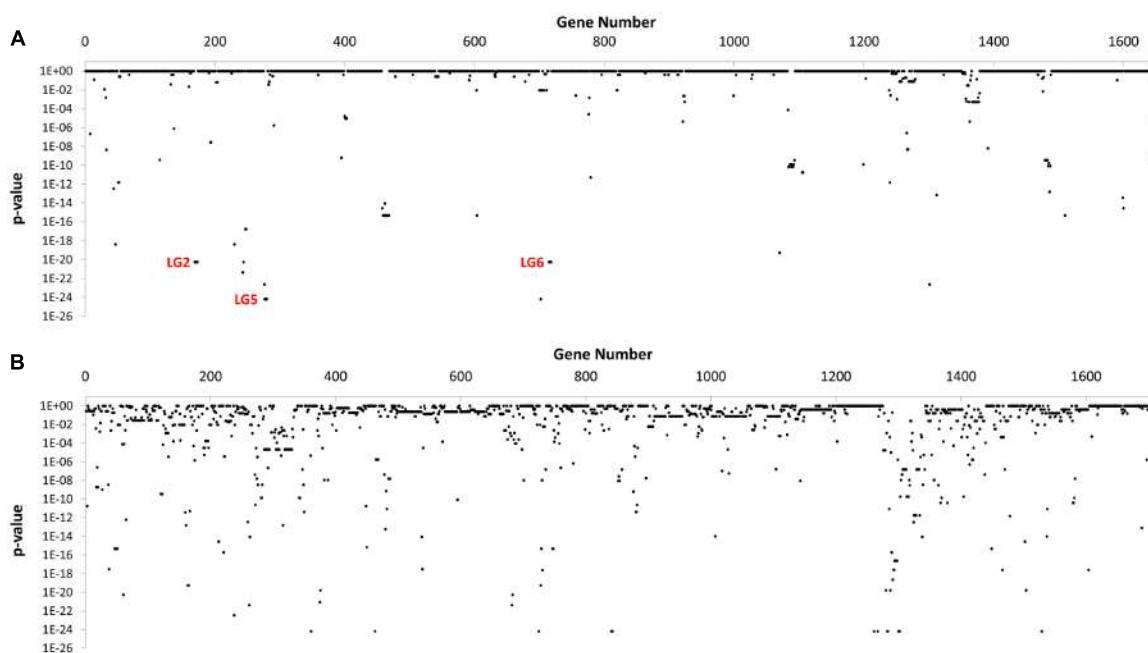


FIGURE 2 | GWAS-based Identification of genes with significant differences in carriage between NCA and CA cohorts. The distribution of *p*-values observed for 3,358 genes in the *C. jejuni* pangenome computed for this study after Genome Fisher analysis of gene carriage data for CA and NCA cohorts. **(A)** Genes from NCTC 11168 genome strains ($n = 1,648$). **(B)** Genes from all other genomes ($n = 1,710$).

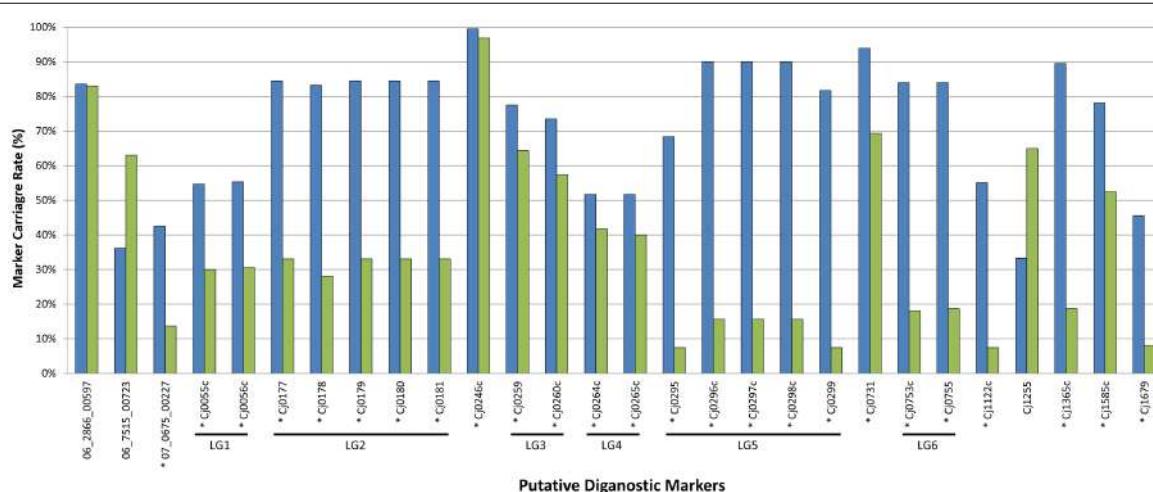


FIGURE 3 | *In silico* validation of putative diagnostic marker genes against expanded CA and NCA cohorts. The putative diagnostic genes identified by GWAS using the original set of 166 genomes were tested for statistical significance with expanded CA (blue bars; $n = 3,742$) and NCA (green bars; $n = 160$) cohorts comprised of additional genomes sequenced in house and from public repositories. Despite the influx of genetically and geographically diverse isolates introduced as part of the expanded dataset, a majority ($n = 25$) of the markers continued to show statistically significant signal with CA subtypes. This suggests that the robust signal detected in the original GWAS analysis stems from genes that appear to have diagnostic value for the identification of *C. jejuni* subtypes with an increased association with campylobacteriosis. *Denotes genes that showed statistically significant signal with CA subtypes.

remainder. A total of 539 CGF subtypes were identified by *in silico* CGF, however, 279 subtypes were novel and had not been previously observed in the C3GFdb and were omitted from the analysis since their epidemiological characteristics could not be determined. Of the remaining 260 CGF subtypes, 38 CGF subtypes (160 genomes) were identified as NCA,

nine CGF subtypes (99 genomes) were identified as UN, and 213 CGF subtypes (3,742 isolates) were identified as CA. Despite the influx of genetically and geographically diverse isolates introduced as part of the expanded dataset, a majority ($n = 25$) of the markers in the original GWAS analysis continued to show statistical significance with CA subtypes;

on average these markers were present in 73% of CA isolates compared to only 36% of NCA isolates (**Figure 3**). Moreover, results of our combinatorial marker analysis show that as few as four markers could be used in combination to detect up to 90% of CA isolates in the validation dataset, with a modest carriage rate of 21% among NCA isolates. These findings suggest that the robust signal detected in the original GWAS analysis stems from genes that appear to have diagnostic value for the identification of *C. jejuni* subtypes with an increased association with campylobacteriosis.

CONCLUSION

A major challenge in the prevention and control of campylobacteriosis is our current inability to identify strains of *C. jejuni* that pose the greatest risk to human health. Addressing this issue would pave the way to better tracking of high-risk strains, leading to a better understanding of their distribution in the food chain and providing critical information towards the development of targeted mitigation strategies to reduce human exposure.

The goal of this study was to identify markers associated with *C. jejuni* lineages known to cause disease in humans and that have a high prevalence in Canada. The genomes of 166 isolates representing 34 highly prevalent *C. jejuni* subtypes were sequenced and a GWAS was performed to identify 28 genes significantly associated with highly prevalent and clinically-related *C. jejuni* subtypes. While some putative gene markers identified as part of this study have previously been associated with important aspects of *C. jejuni* biology including iron acquisition and vitamin B₅ biosynthesis, others represent putative proteins associated with catalysis and transport, which may play roles in processes important for infection and warrant further investigation.

Although these genes were identified within a dataset of Canadian origin, 25 of them continued to display strong statistical significance when validated against a more genetically and geographically diverse dataset. This suggests that they

REFERENCES

- Aickin, M., and Gensler, H. (1996). Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am. J. Public Health* 86, 726–728. doi: 10.2105/AJPH.86.5.726
- Alam, M. T., Petit, R. A., Crispell, E. K., Thornton, T. A., Conneely, K. N., Jiang, Y., et al. (2014). Dissecting vancomycin-intermediate resistance in *Staphylococcus aureus* using genome-wide association. *Genome Biol. Evol.* 6, 1174–1185. doi: 10.1093/gbe/evu092
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421
- Chewapreecha, C., Marttinen, P., Croucher, N. J., Salter, S. J., Harris, S. R., Mather, A. E., et al. (2014). Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet.* 10:e1004547. doi: 10.1371/journal.pgen.1004547
- Clark, C. G., Berry, C., Walker, M., Petkau, A., Barker, D. O. R., Guan, C., et al. (2016). Genomic insights from whole genome sequencing of four clonal outbreak *Campylobacter jejuni* assessed within the global *C. jejuni* population. *BMC Genomics* 17:990. doi: 10.1186/s12864-016-3340-8
- Clark, C. G., Taboada, E., Grant, C. C. R., Blakeston, C., Pollari, F., Marshall, B., et al. (2012). Comparison of molecular typing methods useful for detecting clusters of *Campylobacter jejuni* and *C. coli* isolates through routine surveillance. *J. Clin. Microbiol.* 50, 798–809. doi: 10.1128/JCM.05733-11
- Dasti, J. I., Tareen, A. M., Lugert, R., Zautner, A. E., and Gross, U. (2010). *Campylobacter jejuni*: a brief overview on pathogenicity-associated factors and disease-mediated mechanisms. *Int. J. Med. Microbiol.* 300, 205–211. doi: 10.1016/j.ijmm.2009.07.002
- Duong, T., and Konkel, M. E. (2009). Comparative studies of *Campylobacter jejuni* genomic diversity reveal the importance of core and dispensable genes in the biology of this enigmatic food-borne pathogen. *Curr. Opin. Biotechnol.* 20, 158–165. doi: 10.1016/j.copbio.2009.03.004
- Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., et al. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* 45, 1183–1189. doi: 10.1038/ng.2747

may represent robust markers for clinically-associated *C. jejuni* subtypes, paving the way for future development of molecular assays for rapid identification of *C. jejuni* strains that pose an increased risk to human health.

AUTHOR CONTRIBUTIONS

CB participated in all aspects of laboratory and *in silico* analyses and drafted the manuscript; AW and SM participated in data analysis and drafting of the manuscript; PK, DB, and BH assisted with various aspects of bioinformatics analyses; VG, WA, JT, DI, and ET contributed to study design and writing the manuscript.

FUNDING

Funding for this work was provided by the Alberta Livestock and Meat Association (ALMA) through project 2012F034R, Alberta Innovates Bio Solutions through project BIOFS-12-026, and through the Government of Canada's Genomics Research and Development Initiative.

ACKNOWLEDGMENTS

The authors wish to acknowledge Canada's Michael Smith Genome Sciences Centre, BC, Canada for assistance with sequencing of *C. jejuni* isolates. This work would not have been possible without the collaboration of the many contributors to the Canadian *Campylobacter* Comparative Genomic Fingerprinting database (C3GFdb).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.01224/full#supplementary-material>

- Galanis, E. (2007). *Campylobacter* and bacterial gastroenteritis. *CMAJ* 177, 570–571. doi: 10.1503/cmaj.070660
- Gundogdu, O., Bentley, S. D., Holden, M. T., Parkhill, J., Dorrell, N., and Wren, B. W. (2007). Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. *BMC Genomics* 8:162. doi: 10.1186/1471-2164-8-162
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086
- Havelaar, A. H., van Pelt, W., Ang, C. W., Wagenaar, J. A., van Putten, J. P. M., Gross, U., et al. (2009). Immunity to *Campylobacter*: its role in risk assessment and epidemiology. *Crit. Rev. Microbiol.* 35, 1–22. doi: 10.1080/10408410802636017
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.
- Kim, E.-J., Sabra, W., and Zeng, A.-P. (2003). Iron deficiency leads to inhibition of oxygen transfer and enhanced formation of virulence factors in cultures of *Pseudomonas aeruginosa* PAO1. *Microbiology* 149, 2627–2634. doi: 10.1099/mic.0.26276-0
- Koenraad, P. M. F. J., Rombouts, F. M., and Notermans, S. H. W. (1997). Epidemiological aspects of thermophilic *Campylobacter* in water-related environments: a review. *Water Environ. Res.* 69, 52–63. doi: 10.2175/106143097X125182
- Kruczkiewicz, P., Mutschall, S., Barker, D., Thomas, J. E., Domselaar, G. V. H., Gannon, V. P., et al. (2013). “MIST: a tool for rapid *in silico* generation of molecular data from bacterial genome sequences,” in *Proceedings of Bioinformatics 2013: 4th International Conference on Bioinformatics Models, Methods and Algorithms* (New York, NY: Springer), 316–323.
- Lastovica, A. J., On, S. L., and Zhang, L. (2014). “The family Campylobacteraceae,” in *The Prokaryotes*, eds E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson (Berlin: Springer), 307–335.
- Méric, G., Yahara, K., Mageiros, L., Pascoe, B., Maiden, M. C. J., Jolley, K. A., et al. (2014). A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic *Campylobacter*. *PLoS ONE* 9:e92798. doi: 10.1371/journal.pone.0092798
- Moreno-Hagelsieb, G., and Latimer, K. (2008). Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 24, 319–324. doi: 10.1093/bioinformatics/btm585
- Muley, V. Y., and Acharya, V. (2013). “Chromosomal proximity of genes as an indicator of functional linkage,” in *Genome-Wide Prediction and Analysis of Protein-Protein Functional Linkages in Bacteria*, eds V. Y. Muley and V. Acharya (New York, NY: Springer), 33–42. doi: 10.1007/978-1-4614-4705-4_4
- Nachamkin, I. (2002). Chronic effects of *Campylobacter* infection. *Microbes Infect.* 4, 399–403. doi: 10.1016/S1286-4579(02)01553-8
- Nachamkin, I., Allos, B. M., and Ho, T. (1998). *Campylobacter* species and Guillain-Barré syndrome. *Clin. Microbiol. Rev.* 11, 555–567.
- Naikare, H., Butcher, J., Flint, A., Xu, J., Raymond, K. N., and Stintzi, A. (2013). *Campylobacter jejuni* ferric-enterobactin receptor CfrA is TonB3 dependent and mediates iron acquisition from structurally different catechol siderophores. *Metalomics* 5, 988–996. doi: 10.1039/C3MT20254B
- Naikare, H., Palyada, K., Panciera, R., Marlow, D., and Stintzi, A. (2006). Major role for FeoB in *Campylobacter jejuni* ferrous iron acquisition, gut colonization, and intracellular survival. *Infect. Immun.* 74, 5433–5444. doi: 10.1128/IAI.00052-06
- Palyada, K., Threadgill, D., and Stintzi, A. (2004). Iron acquisition and regulation in *Campylobacter jejuni*. *J. Bacteriol.* 186, 4714–4729. doi: 10.1128/JB.186.14.4714-4729.2004
- Parkhill, J., Wren, B. W., Mungall, K., Ketley, J. M., Churcher, C., Basham, D., et al. (2000). The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* 403, 665–668. doi: 10.1038/35001088
- Pintar, K. D. M., Thomas, K. M., Christidis, T., Otten, A., Nesbitt, A., Marshall, B., et al. (2016). A Comparative exposure assessment of *Campylobacter* in Ontario, Canada. *Risk Anal.* 37, 677–715. doi: 10.1111/risa.12653
- Read, T. D., and Massey, R. C. (2014). Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med.* 6:109. doi: 10.1186/s13073-014-0109-z
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153
- Sheppard, S. K., Didelot, X., Méric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., et al. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11923–11927. doi: 10.1073/pnas.1305559110
- Silva, J., Leite, D., Fernandes, M., Mena, C., Gibbs, P. A., and Teixeira, P. (2011). *Campylobacter* spp. as a foodborne pathogen: a review. *Front. Microbiol.* 2:200. doi: 10.3389/fmicb.2011.00200
- Suzuki, H., and Yamamoto, S. (2009). *Campylobacter* contamination in retail poultry meats and by-products in the world: a literature survey. *J. Vet. Med. Sci.* 71, 255–261. doi: 10.1292/jvms.71.255
- Taboada, E. N., Ross, S. L., Mutschall, S. K., Mackinnon, J. M., Roberts, M. J., Buchanan, C. J., et al. (2012). Development and validation of a comparative genomic fingerprinting method for high-resolution genotyping of *Campylobacter jejuni*. *J. Clin. Microbiol.* 50, 788–797. doi: 10.1128/JCM.00669-11
- Taboada, E. N., van Belkum, A., Yuki, N., Acedillo, R. R., Godschalk, P. C., Koga, M., et al. (2007). Comparative genomic analysis of *Campylobacter jejuni* associated with Guillain-Barré and Miller Fisher syndromes: neuropathogenic and enteritis-associated isolates can share high levels of genomic similarity. *BMC Genomics* 8:359. doi: 10.1186/1471-2164-8-359
- Thépault, A., Méric, G., Rivoal, K., Pascoe, B., Mageiros, L., Touzain, F., et al. (2017). Genome-wide identification of host-segregating epidemiological markers for source attribution in *Campylobacter jejuni*. *Appl. Environ. Microbiol.* 83:e3085-16. doi: 10.1128/AEM.03085-16
- Thomas, M. K., Murray, R., Flockhart, L., Pintar, K., Pollari, F., Fazil, A., et al. (2013). Estimates of the burden of foodborne illness in Canada for 30 specified pathogens and unspecified agents, circa 2006. *Foodborne Pathog. Dis.* 10, 639–648. doi: 10.1089/fpd.2012.1389
- Ward, N., and Moreno-Hagelsieb, G. (2014). Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: How much do we miss? *PLoS ONE* 9:e101850. doi: 10.1371/journal.pone.0101850
- Whiley, H., van den Akker, B., Giglio, S., and Bentham, R. (2013). The role of environmental reservoirs in human campylobacteriosis. *Int. J. Environ. Res. Public Health* 10, 5886–5907. doi: 10.3390/ijerph10115886
- Williams, A., and Oyarzabal, O. A. (2012). Prevalence of *Campylobacter* spp. in skinless, boneless retail broiler meat from 2005 through 2011 in Alabama, USA. *BMC Microbiol.* 12:184. doi: 10.1186/1471-2180-12-184
- Yahara, K., Méric, G., Taylor, A. J., de Vries, S. P. W., Murray, S., Pascoe, B., et al. (2016). Genome-wide association of functional traits linked with *Campylobacter jejuni* survival from farm to fork. *Environ. Microbiol.* 19, 361–380. doi: 10.1111/1462-2920.13628

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Buchanan, Webb, Mutschall, Kruczkiewicz, Barker, Hetman, Gannon, Abbott, Thomas, Inglis and Taboada. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Evolution and Diversity of *Listeria monocytogenes* from Clinical and Food Samples in Shanghai, China

Jianmin Zhang^{1†}, Guojie Cao^{2†}, Xuebin Xu³, Marc Allard⁴, Peng Li⁵, Eric Brown⁴, Xiaowei Yang⁶, Haijian Pan⁶ and Jianghong Meng^{2*}

¹ National and Regional Joint Engineering Laboratory for Medicament of Zoonosis Prevention and Control, Key Laboratory of Zoonosis Prevention and Control of Guangdong Province, College of Veterinary Medicine, South China Agricultural University, Guangzhou, China, ² Department of Nutrition and Food Science and Joint Institute for Food Safety and Applied Nutrition, University of Maryland, College Park, College Park, MD, USA, ³ Shanghai Municipal Center for Disease Control and Prevention, Shanghai, China, ⁴ Division of Microbiology, Office of Regulatory Science, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, MD, USA, ⁵ Institute of Disease Control and Prevention, Academy of Military Medical Science, Beijing, China, ⁶ Department of Food Science & Technology, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China

OPEN ACCESS

Edited by:

Jennifer Ronholm,
Health Canada, Canada

Reviewed by:

Jinshui Zheng,

Huazhong Agricultural University,

China

Min Yue,

University of Pennsylvania, USA

*Correspondence:

Jianghong Meng
jmeng@umd.edu

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 06 May 2016

Accepted: 07 July 2016

Published: 22 July 2016

Citation:

Zhang J, Cao G, Xu X, Allard M, Li P, Brown E, Yang X, Pan H and Meng J (2016) Evolution and Diversity of *Listeria monocytogenes* from Clinical and Food Samples in Shanghai, China. *Front. Microbiol.* 7:1138.
doi: 10.3389/fmicb.2016.01138

Listeria monocytogenes is a significant foodborne pathogen causing severe systemic infections in humans with high mortality rates. The objectives of this work were to establish a phylogenetic framework of *L. monocytogenes* from China and to investigate sequence diversity among different serotypes. We selected 17 *L. monocytogenes* strains recovered from patients and foods in China representing serotypes 1/2a, 1/2b, and 1/2c. Draft genome sequences were determined using Illumina MiSeq technique and associated protocols. Open reading frames were assigned using prokaryotic genome annotation pipeline by NCBI. Twenty-four published genomes were included for comparative genomic and phylogenetic analysis. More than 154,000 single nucleotide polymorphisms (SNPs) were identified from multiple genome alignment and used to reconstruct maximum likelihood phylogenetic tree. The 41 genomes were differentiated into lineages I and II, which consisted of 4 and 11 subgroups, respectively. A clinical strain from China (SHL009) contained significant SNP differences compared to the rest genomes, whereas clinical strain SHL001 shared most recent common ancestor with strain SHL017 from food. Moreover, clinical strains SHL004 and SHL015 clustered together with two strains (08-5578 and 08-5923) recovered from an outbreak in Canada. Partial sequences of a plasmid found in the Canadian strain were also present in SHL004. We investigated the presence of various genes and gene clusters associated with virulence and subgroup-specific genes, including internalins, *L. monocytogenes* pathogenicity islands (LPIs), *L. monocytogenes* genomic islands (LGIs), stress survival islet 1 (SSI-1), and clustered regularly interspaced short palindromic repeats (CRISPR)/cas system. A novel genomic island, denoted as LGI-2 was identified. Comparative sequence analysis revealed differences among the *L. monocytogenes* strains related to virulence, survival abilities, and attributes against foreign genetic elements. *L. monocytogenes* from China were genetically diverse. Strains from clinical specimens and food related closely suggesting foodborne transmission of human listeriosis.

Keywords: *L. monocytogenes*, evolution, whole genome analysis, plasmid, China

INTRODUCTION

Listeria monocytogenes causes severe systemic infections with high mortality rates (Toledo-Arana et al., 2009; Kuenne et al., 2013) and has been responsible for numerous outbreaks in North America and Europe during recent years (Gilmour et al., 2010; Jackson et al., 2011; Smith et al., 2011; Laksanalamai et al., 2012; Schoder et al., 2014). *L. monocytogenes* is able to survive and grow under a wide range of temperature and pH conditions with a significant tolerance to salt (Gilmour et al., 2010). This ubiquitous pathogen imposes a risk with significant economic burden to public health (Toledo-Arana et al., 2009).

L. monocytogenes consists of four evolutionary lineages and 13 identified serotypes, with serotypes 1/2a, 1/2b, 1/2c, and 4b causing most human listeriosis cases (Swaminathan and Gerner-Smidt, 2007; Ragon et al., 2008; den Bakker et al., 2013). Serotypes 1/2b and 4b belong to lineage I; 1/2a and 1/2c belong to lineage II (Ragon et al., 2008). Serotypes 1/2a, 1/2b, and 1/2c accounted for more than 90% of *L. monocytogenes* isolated from food in China (Wang et al., 2012).

Central to *L. monocytogenes* pathogenesis is the ability to invade and cross host barriers (Bergmann et al., 2013) and to secrete proteins beyond the cell surface using internalins and various secretion systems (Desvaux and Hebraud, 2006). The internalin family is composed of important proteins involving virulence activities of *L. monocytogenes*. Internalins A (InlA) and B (InlB) played essential roles in invasion activities (Dussurget et al., 2004). Several additional internalins InlJ, InlI, and InlK have been identified recently (Sabet et al., 2005; Neves et al., 2013; Becavin et al., 2014). There are six protein secretion systems in *L. monocytogenes*, including Sec (secretion system), Tat (twin-arginine translocation), FPE (fimbriae protein exporter), FEA (flagella export apparatus), holins, and Wss (WXG100 secretion system) (Desvaux and Hebraud, 2006).

Similar to *Salmonella* (Cao et al., 2013) and *Escherichia coli* (Ju et al., 2014), *L. monocytogenes* contains genomic islands playing important roles in virulence, such as *Listeria* pathogenicity islands (LIPIs) (Gonzalez-Zorn et al., 2000; Clayton et al., 2014) and *Listeria* genomic islands (LGIs) (Gilmour et al., 2010). LIPI-1 contains virulence determinants, like *hly*, *plcAB*, and *actA* (Gonzalez-Zorn et al., 2000). Moreover, *L. monocytogenes* carries a gene cluster termed stress survival islet 1 (SSI-1), which is composed of five genes contributing to the survival of cells in suboptimal conditions of food environments such as low pH and high salt concentrations (Ryan et al., 2010).

Clustered regularly interspaced short palindromic repeats (CRISPR)/cas system is considered a bacterial immune system against invading genetic fragments by targeting specific sequence including phages and plasmids (Touchon and Rocha, 2010). The presence of functional CRISPR/cas system has a negative correlation with resistance to antibiotics in *Enterococci* (Palmer and Gilmore, 2010). In addition, non-coding RNAs regulate gene expression through hybridizing with mRNA or binding to proteins to modulate their activities under different conditions (Kuenne et al., 2013). Both of these systems contribute to the pathogenicity, antimicrobial resistance, and metabolism of *L. monocytogenes*.

The objectives of the current study were to provide a phylogenetic framework of human and foodborne *L. monocytogenes* from China, and to reveal genomic sequence diversities of *L. monocytogenes*. The data should assist in a better understanding of the evolution and genetic diversity of *L. monocytogenes*.

MATERIALS AND METHODS

Bacterial Strains

To determine the genetic diversity and to identify the genetic characteristics of *L. monocytogenes*, 12 clinical strains belonging to different PFGE profiles were selected among 50 strains from Shanghai, China (2004–2012) (unpublished data) (Table 1). Moreover, five strains of *L. monocytogenes* from food were included to investigate a possible connection of food to the clinical cases. The strains represented serotypes 1/2a ($n = 8$), 1/2b ($n = 5$), and 1/2c ($n = 4$).

Genome Sequencing and Annotation

Whole genome sequencing was performed on the 17 strains using Illumina MiSeq technique and associated protocols (Illumina, San Diego, CA) with MiSeq Reagent Kit v2 (500 cycle), Nextera XT DNA Sample Preparation kit, and Nextera XT Index Kit. Draft genome data were assembled using CLC Genomic Workbench *de novo* and were annotated by NCBI using Prokaryotic Genomes Annotation Pipeline (Klimke et al., 2009). These draft genomes were deposited in GenBank under the following accession numbers: SHL001 (APIB00000000), SHL002 (APIB00000000), SHL004 (APIB00000000), SHL005 (APIB00000000), SHL006 (APIB00000000), SHL007 (APIB00000000), SHL008 (APIB00000000), SHL009 (APIB00000000), SHL010 (APIB00000000), SHL011 (APIB00000000), SHL012 (APIB00000000), SHL013 (APIB00000000), SHL014 (AWWQ00000000), SHL015 (AWWR00000000), SHL016 (AWWS00000000), SHL017 (AWWT00000000), and SHL018 (AWWU00000000).

Genomic Analysis

In addition to the 17 genomes, 24 publicly available *L. monocytogenes* genomes were selected to determine the evolutionary relationship among *L. monocytogenes* (Table 2). Single nucleotide polymorphisms (SNPs) were identified based on core genome alignments using progressive Mauve (Darling et al., 2010) and customized in-house script. Clonal complexes were determined by multilocus sequence typing (MLST) analysis using seven housekeeping genes, including *abcZ*, *bgIA*, *cat*, *dapE*, *dat*, *Idh*, and *IhkA* (Cantinelli et al., 2013). To reconstruct evolutionary relatedness among the genomes, we used Genetic Algorithm for Rapid Likelihood Inference (GARLI 2.0) to perform a maximum likelihood analysis (Zwickl, 2006) with 1000 bootstrap replicates and GTR+I+G nucleotide substitution model based on the SNPs we identified. Pairwise distance matrix with the number of nucleotide differences was calculated using MEGA 5.10 (Tamura et al., 2011) with 10,000 bootstrap replications. CRISPR arrays and cas gene clusters were identified using CRISPRFinder (Grissa et al., 2007). Stand-alone blast (blast 2.27+) (Altschul et al., 1990) was used to

TABLE 1 | Metadata Associated with 17 *L. monocytogenes* strains from Shanghai, China.

Strain	Serotype	Lineage	Year	Sources	Contigs	N50 size	STs
SHL001	1/2a	II	2007	Human#	33	476,844	381
SHL004	1/2a	II	2008	Human	18	579,300	8
SHL005	1/2a	II	2008	Human	17	437,049	7
SHL009	1/2a	II	2012	Human*	32	541,739	91
SHL011	1/2a	II	2011	Human	17	543,519	29
SHL013	1/2a	II	2012	Human*	20	358,858	391
SHL002	1/2b	I	2007	Human	22	476,844	3
SHL007	1/2b	I	2011	Human	43	355,359	87
SHL008	1/2b	I	2012	Human	26	293,078	3
SHL010	1/2b	I	2012	Human	84	259,950	2
SHL012	1/2b	I	2010	Human#	23	355,398	87
SHL006	1/2c	II	2010	Human	30	476,139	9
SHL015	1/2a	II	2008	Beef	15	425,029	8
SHL017	1/2a	II	2004	Bean	17	726,747	381
SHL014	1/2c	II	2008	Pork	23	477,674	9
SHL016	1/2c	II	2008	Fish	17	512,641	9
SHL018	1/2c	II	2004	Vegetables	18	429,471	9

#Cerebrospinal fluid; all other clinical strains were isolated from blood.

*The host of these strains died.

determine the presence/absence of 117 virulence related genes, 150 non-coding RNAs (Izar et al., 2011), and other gene clusters. Strain specific elements within subgroups were identified using cluster_smallmem command in the USEARCH package (Edgar, 2010) with 90% identities as threshold. The genome organization comparison for subgroup IIk was displayed using BRIG 0.95 with 90 and 70% as upper and lower identity threshold, respectively (Alikhan et al., 2011).

RESULTS

Our findings revealed the phylogeny of 41 *L. monocytogenes* strains from diverse sources and geographic locations, and provided insights into their sequence diversities. The size of draft genomes of the 17 strains from Shanghai, China (Table 1) ranged from 2.86 Mb (SHL013) to 3.12 Mb (SHL002) (Table 2). SHL013 contained the lowest number of genes ($n = 2821$) whereas SHL002 had the highest number of genes ($n = 3113$). The average genome size in lineages I and II was 2.93 Mb and 2.98 Mb, respectively.

Phylogenetic Analysis between Strains

A maximum likelihood (ML) phylogenetic tree was constructed using more than 154,000 SNPs, which were identified from core genome alignments (Figure 1). The 41 *L. monocytogenes* genomes were divided into two lineages based on phylogenetic data. Serotype 1/2b strains belonged to lineage I, whereas serotypes 1/2a and 1/2c strains belonged to lineage II. These two lineages were further split into 4 (Ia to Id) and 11 (IIa to IIk) subgroups, respectively (Figure 1). The number of SNPs differences (standard deviation) between the four subgroups of lineage I ranged from 7593 (± 47 SNPs) to 10,681 SNPs (± 94

SNPs) (Table S1). The number of SNPs differences between lineage II subgroups were more than 18,692 SNPs (± 86 SNPs) with one exception (Table S1). The SNPs differences between subgroups IIIi (strain EGDe) and IIj (serotype 1/2c) were only 1112 SNPs (± 30 SNPs). Serotype 1/2c strains originated from EGDe, which was the middle point between serotypes 1/2a and 1/2c. EGDe and all serotype 1/2 strains belonged to clonal complex (CC) 9.

In lineage I, some clinical strains were closely related each other (Figure 1, Table S2). Strain SHL007 (2011) displayed only 30 SNPs (± 5 SNPs) differences from SHL012 (2010), which belonged to sequence type (ST) 87 same as reference genome FSL J1-175. The SNPs difference between SHL002 (2007) and SHL008 (2012) was 194 SNPs (± 8 SNPs), and both strains belonged to ST3.

Lineage II strains displayed a divergent structure between subgroups (Table S1). Eight 1/2a strains from China were scattered in different subgroups with large SNPs differences. SHL013 recovered from an infant who died of listeriosis had more than 24,000 SNPs differences compared to other genomes except FSL J2-003. Another lethal strain, SHL009, also showed more than 20,000 SNPs differences compared to the rest genomes. Several foodborne and clinical strains were genetically closed. However, there was no epidemiological data available to make any foodborne illness connection. For example, the difference was 60 SNPs (± 4 SNPs) between SHL001 (human) and SHL017 (bean), and only 11 SNPs (± 2 SNPs) between SHL004 (human), and SHL015 (beef) (Table S2). Serotype 1/2c strains showed a clonal structure and the differences between the 1/2c genomes were no more than 300 SNPs (Table S2).

Based on SNPs difference, several strains from China appeared to have a close evolutionary relationship to those from North

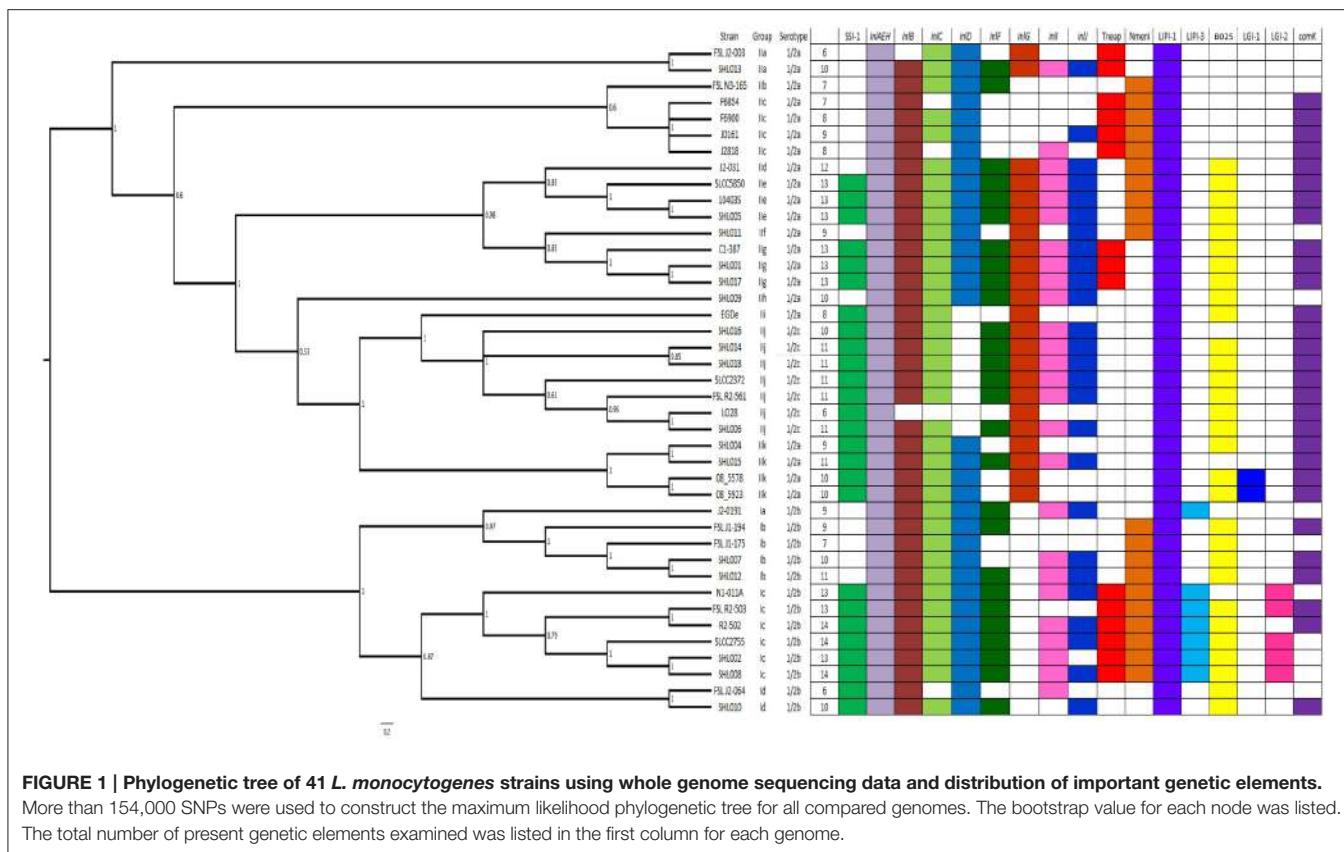
TABLE 2 | Sequencing Statistics for 41 Selected Strains of *L. monocytogenes*.

Strain	Serotype/Lineage	Genome Size	GC Content	CDS	Source	Year	Country	GenBank Accession
SHL001	1/2a, II	2.95	37.9	2964	Human	2007	China	APIB00000000
SHL004	1/2a, II	3.01	37.8	3018	Human	2008	China	APID00000000
SHL005	1/2a, II	2.88	37.9	2859	Human	2008	China	APIE00000000
SHL009	1/2a, II	2.87	37.9	2866	Human	2012	China	APII00000000
SHL011	1/2a, II	2.87	37.9	2847	Human	2011	China	APIK00000000
SHL013	1/2a, II	2.86	37.9	2821	Human	2012	China	APIM00000000
SHL015	1/2a, II	2.96	38.0	2946	Beef	2008	China	AWWIR00000000
SHL017	1/2a, II	2.95	38.0	2937	Bean	2004	China	AWWT00000000
SHL002	1/2b, I	3.12	37.9	3113	Human	2007	China	APIC00000000
SHL007	1/2b, I	2.98	37.9	2995	Human	2011	China	APIG00000000
SHL008	1/2b, I	3.01	37.9	2990	Human	2012	China	APIH00000000
SHL010	1/2b, I	3.08	37.9	3112	Human	2012	China	APIJ00000000
SHL012	1/2b, I	2.93	37.9	2906	Human	2010	China	APIL00000000
SHL006	1/2c, II	2.93	37.9	2959	Human	2010	China	APIF00000000
SHL014	1/2c, II	2.95	37.9	2952	Pork	2008	China	AWWQ00000000
SHL016	1/2c, II	2.97	37.7	2992	Fish	2008	China	AWWS00000000
SHL018	1/2c, II	2.94	37.8	2948	Vegetables	2004	China	AWWU00000000
10403S	1/2a, II	2.90	38.0	2814	Human	1968	U.S.	CP002002
F6900	1/2a, II	2.97	37.7	3005	Human	1989	U.S.	AARU02000000
J2-031	1/2a, II	2.96	37.9	2924	Human	1996	U.S.	CP006593
J2818	1/2a, II	2.97	37.7	3083	Human	2000	U.S.	AARX02000000
J0161	1/2a, II	3.00	37.9	2955	Human	2000	U.S.	CP002001
08-5578	1/2a, II	3.03	38.0	3088	Human	2008	Canada	CP001602.1
08-5923	1/2a, II	3.00	38.0	2966	Human	2008	Canada	CP001604
SLCC5850	1/2a, II	2.91	38.0	2866	Rabbit	1924	UK	FR733647
EGD-e	1/2a, II	2.94	38.0	2846	Rabbit	1926	UK	AL591824.1
F6854	1/2a, II	2.95	37.8	2967	Hot dog	1988	U.S.	AADQ01000000
C1-387	1/2a, II	2.99	37.9	2953	Food	1999	U.S.	CP006591
FSL J2-003	1/2a, II	2.74	37.8	2937	N.a.	N.A.	U.S.	AARM02000000
FSL N3-165	1/2a, II	2.88	37.8	2890	Soil	N.A.	U.S.	AARQ02000000
FSL R2-503	1/2b, I	2.99	37.8	3027	Human	1994	U.S.	AARR00000000
SLCC2755	1/2b, I	2.97	38.1	2940	Human	N.A.	N.A.	NC_018587
FSL J1-194	1/2b, I	2.99	37.8	3012	Human	N.A.	U.S.	AARJ00000000
R2-502	1/2b, I	3.03	37.9	2984	Food	1994	U.S.	CP006594
J2-1091	1/2b, I	2.98	37.9	2912	Animal	1995	U.S.	CP006596
FSL J1-175	1/2b, I	2.87	37.9	3147	Water	2006	U.S.	AARK00000000
FSL J2-064	1/2b, I	2.83	37.9	2934	Food	N.A.	N.A.	AARO00000000
N1-011A	1/2b, I	3.01	37.9	3059	Environment	N.A.	U.S.	CP006597
LO28	1/2c, II	2.68	37.8	2999	Human	N.A.	N.A.	AARY00000000
FSL R2-561	1/2c, II	2.97	38.0	2910	N.A.	N.A.	N.A.	NC_017546
SLCC 2372	1/2c, II	2.97	38.0	2990	N.A.	N.A.	N.A.	NC_018588

America. SHL007 to SHL008 had no more than 120 SNPs difference compared to FSLJ1-175 and R2-502 from the United States (**Table S2**). Clinical strain SHL004 displayed 94 SNPs (± 7 SNPs) and 93 SNPs (± 7 SNPs) differences compared to strains 08-5578 and 08-5923, respectively, which were recovered from a large foodborne outbreak in Canada in 2008 (**Table S1**). Similarly, SHL015 had only 97 SNPs (± 7 SNPs) and 96 SNPs (± 6 SNPs) differences compared to 08-5578 and 08-5923, respectively.

Phylogenetic and Comparative Genomic Analyses between Strains from China (SHL004 and SHL015) and the Outbreak Strains (08-5578 and 08-5923) of Canada

SHL004 and SHL015 were closely clustered with 08-5578 and 08-5923 in subgroup IIk (**Figure 1**). They belonged to CC8 that was identified as epidemic clone V (ECV). The SNPs differences between these genomes ranged from 11 to 97 SNPs (**Table S2**).



Their genome organizations were displayed in **Figure S1**. We also identified plasmids sequences in SHL004 and SHL015 (**Figure S2**). Contig number 9 of SHL015 (86,633 bp, GC content: 37%, 92 ORFs) showed 99% identities and 100% cover compared to plasmid pLMR479a (86,652 bp, accession number: HG813248). Contig number 9 of SHL004 (79,013 bp, GC content: 36.7%, 82 ORFs) showed 100% identities compared to pLMR479a. Both plasmid sequences from SHL004 and SHL015 were also highly conserved compared to plasmid pLM5578 (77,054 bp, accession number: CP001603) from strain 08-5578.

There were certain variations in virulence determinants among the four genomes. SHL004 and SHL015 did not contain LGI-1 whereas 08-5578 and 08-5923 did. They contained one gene cluster encoding prophage proteins (phage B025) except SHL015. SHL015 carried genes *inlF*, *inlI*, and *inlJ*, which were absent in the other three strains. An evolutionary model for subgroup IIk strains was proposed (**Figure 2**), in which the last common ancestor containing plasmid was divided into the Chinese and Canadian lineages (Gilmour et al., 2010).

Distribution of Important Genetic Elements

Of 117 genes related to virulence, metabolism, and regulations determined (**Table S3**), 101 genes were present in all 41 genomes. We also found genes were present only in certain subgroups. For example, *lmo0150* (hypothetical protein) was present in subgroups IIa, IIi, IIj, and IIk; *lmo0471* (hypothetical protein) was found in subgroups IIi and IIj. Additionally, we identified strain

specific genes within subgroups via comparing the genomes in the same subgroup (**Table S4**). The distribution of other genetic elements follows.

Stress Survival Islet (SSI-1)

SSI-1 was present in both lineages (Ic, Id, IIe, IIg, IIIi, IIj, and IIk; **Figure 1**). The common ancestor of subgroups Ic and Id may have acquired SSI-1. The same happened to subgroups IIIi, IIj, and IIk. In contrast, it seems that the ancestors of subgroups IIe and IIg obtained SSI-1 independently. A gene encoding transcriptional regulator was identified upstream of SSI-1. There was no other inserted sequence found in those without SSI-1 (Hein et al., 2011).

Internalins

The number of internalins in the genomes examined ranged from 5 (F6854 and LO28) to 10 (SHL013, SHL015, and subgroups IIe, IIg, IIh; **Figure 1**). A gene cluster encoded four internalins including *inlG*, *inlH* (*inlC2*), *inlD*, and *inlE* (5' to 3'). *InlH* and *inlE* were present in all genomes whereas *inlG* and *inlD* were found in different subgroups (**Figure 1**). *InlG* was present in the lineage II subgroups except IIb and IIc. However, lineage I strains did not contain *inlG*. Gene *inlD* was present in both lineages but subgroups IIIi and IIj. SHL009 and SHL013 carried all four genes. Additionally, the presence of *inlC*, *inlF*, *inlI*, and *inlJ* were inconsistently scattered between different subgroups (**Figure 1**). SHL015 possessed all four internalins;

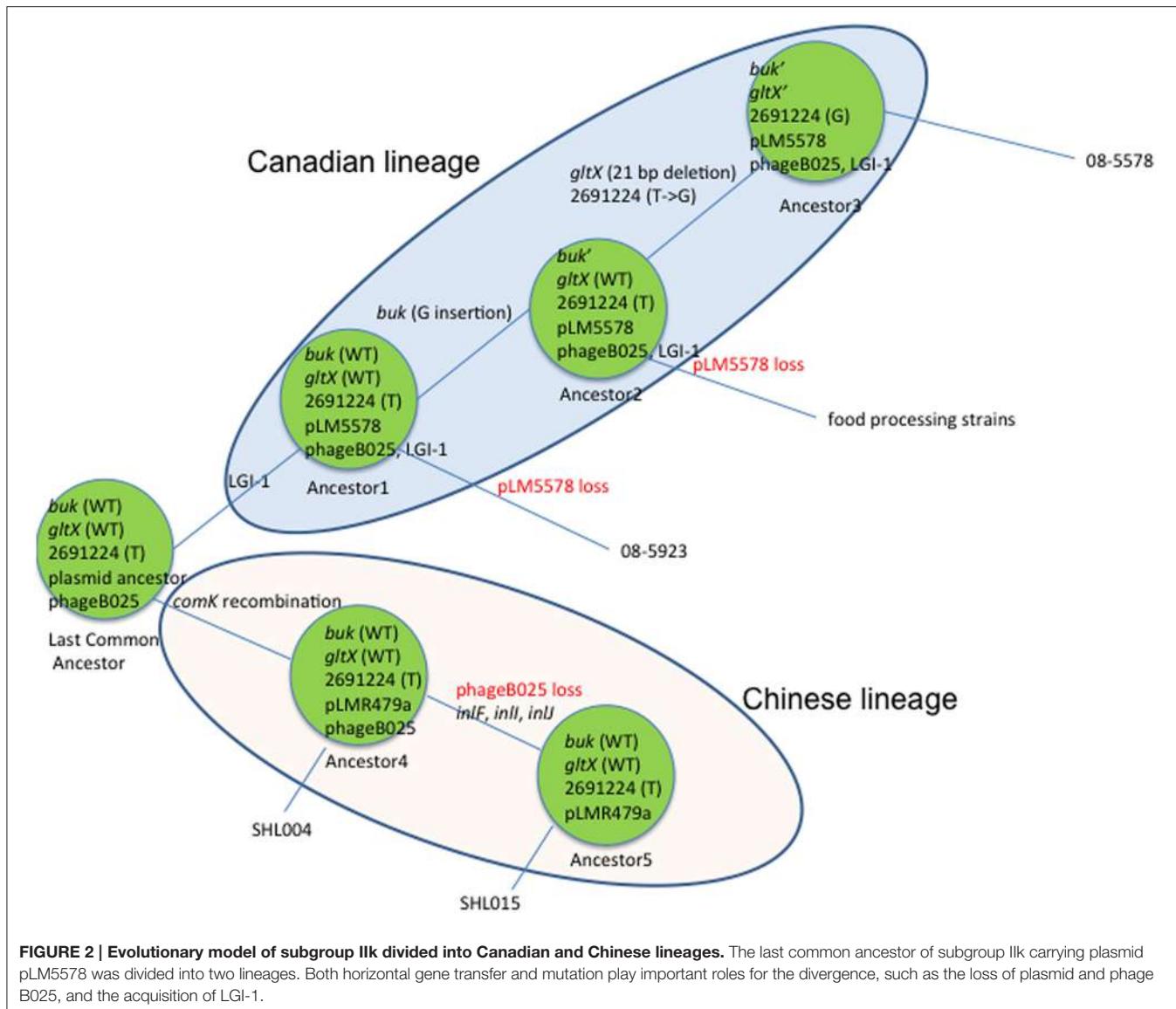


FIGURE 2 | Evolutionary model of subgroup IIk divided into Canadian and Chinese lineages. The last common ancestor of subgroup IIk carrying plasmid pLM5578 was divided into two lineages. Both horizontal gene transfer and mutation play important roles for the divergence, such as the loss of plasmid and phage B025, and the acquisition of LGI-1.

however, SHL004 contained only *inlC*, which appeared to be lost independently four times on the phylogeny (Figure 1).

Non-coding RNAs (ncRNAs)

Among 150 ncRNAs (Table S5), 109 ncRNAs were present in all genomes with truncated lengths. For example, *rli91* in F6854 was 42 out of 88 bp. The rest ncRNAs remained in certain subgroups/strains (Table S5). *Rli29*, *rli85*, and *anti0469* were acquired by the common ancestor of subgroups IIi and IIj. *RliC* was identified in subgroups Id, IIi, and IIj. *Rli110* existed in all genomes except subgroup IIa genomes. *anti1457*, *anti1749*, *anti1758*, and *anti1974* were found in lineage II. *Rli28*, *rli50*, *rli78*, and *rli112* were identified in subgroups IIc, IIg, IIi, IIj, and IIk.

Secretion System Proteins

Secretory pathway components of six identified secretion systems were present in all genomes except that certain genes were absent

in some genomes (Table S6). The *tatA* gene was found exclusively in lineage II whereas *tatC* was present in lineage II and subgroup Id. *EsaC* was found in subgroups Ia, Ib, Id, IIIi, and IIj. *SecE* was identified in all strains but F6854. *LspB* was only identified in EGDe.

Listeria Pathogenicity Islands (LIPIs), Listeria Genomic Islands (LGIs), and Prophage

The 41 genomes all contained LIPI-1 including *prfA*, *plcA*, *plcB*, *hly*, *mpl*, and *actA*. A 37-kb gene cluster encoding prophage proteins was located at the same location of LIPI-2 in *Listeria* (Figure 1, Table S7). This gene cluster was identified as phage B025 and contained no genes associated with virulence or antimicrobial resistance. B025 may have been acquired by the common ancestor of subgroups IId, IIe, IIf, and IIg. Independent loss events of B025 happened in subgroups IIj, and IIk (Figure 1). Moreover, LIPI-3 may have been independently obtained in

subgroups Ia and Ic, encoding *llsG*, *llsH*, *llsX*, *llsB*, *llsY*, *llsD*, and *llsP* (Clayton et al., 2014).

LGI-1 exclusively existed in 08-5578 and 08-5923 in subgroup IIk. An insertion encoding 50 genes was identified at the same locus of LGI-1 in subgroup Ic except R2-502, termed as LGI-2 (Table S8). LIG-2 originated from phage and did not contain virulence genes based on the current annotation.

ComK Prophage Junction Fragment

The *comK* phage insertions were present in 12 strains from China (Figure 1). These insertions carried divergent components with both indels and substitutions. Some strains in the same subgroup contained identical *comK* phage insertions, such as SHL001 and SHL017, SHL004, and SHL015. In contrast, subgroup Ib strains SHL007 and SHL012 carried different components in the *comK* junction fragment.

CRISPR/cas System

Some *L. monocytogenes* strains from the same subgroup contained identical CRISPR spacer arrays, such as subgroup Id (Table S9). EGDe (1/2a, subgroup III) contained the same spacers as 1/2c strains (subgroup IIj). Several strains from China contained identical spacers that were different from those of the rest strains in the same subgroup, such as subgroup Ic strains SHL002 and SHL008 (Table S9). The number of spacers in different strains varied. There were 68 spacers in SHL001 and SHL017, but no spacer present in SHL009.

Two *cas* gene clusters were identified, Tneap and Nmeni (Figure 1). Tneap consisted of *cas6*, *cst1*, *cst2*, *cas5t*, *cas3*, *cas1*, and *cas2-1*, whereas Nmeni included *csn2*, *cas2-2*, *cas1*, and *csn1*. Subgroups Ic and Ic strains carried both *cas* clusters. In contrast, subgroups Ia, Id, IIh, III, IIj, and IIk did not contain any *cas* genes. The other subgroups contained either Tneap or Nmeni.

DISCUSSION

In the present study, we sequenced 17 *L. monocytogenes* strains recovered from humans and food in China and built a phylogenetic framework for *L. monocytogenes* using these genomes against publicly available data from diverse sources and locations. *L. monocytogenes* serotypes 1/2a and 1/2b displayed a divergent structure. In contrast, serotype 1/2c showed a clonal structure with small SNPs differences (Figure 1, Table S2). The strains from China contained extensive diversification, suggesting the evolutionary diversity and complex of *L. monocytogenes* in China.

Our findings provided detailed and comprehensive information on *L. monocytogenes* evolution and diversity. Four genomes belonging to CC8 were grouped together, SHL004 (human), SHL015 (beef), and two outbreak isolates 08-5578 and 08-5923 in Canada (Gilmour et al., 2010). Intriguingly, the isolation times for these two pair strains were close, September and August 2008, respectively. The proposed evolutionary model (Figure 2) suggests this *L. monocytogenes* clone complex may have circulated globally and the strains from different geographic locations have split into two subgroups.

Clinical strains of serotype 1/2b recovered from different years in Shanghai, China, (SHL002 in 2007 and SHL008 in 2012; SHL007 in 2011 and SHL012 in 2010) showed close genetic relatedness, suggesting these clones remain and circulate locally. SHL002 and SHL008 shared most recent common ancestor. SHL007 and SHL012 also originated from the same ancestor. Additionally, clinical strains of serotype 1/2a were grouped together with those from foods (SHL015 from beef and SHL017 from bean). Such an association indicated a possible foodborne transmission of listeriosis to humans.

The distribution of virulence factors indicated variations in virulence potentials and evolutionary histories in different subgroups. The mean numbers of virulence factors examined in subgroups Ic and IIg (Figure 1) were higher than other subgroups suggesting a link between these genetic factors and observed differences in pathogenicity, survival, and risk in causing diseases.

Internalins play essential roles in host-cell interactions of *L. monocytogenes* (Bierne et al., 2007) and the presence and distribution of internalins relate to differences in virulence potential (Rychli et al., 2014). *InlC* is essential in liver infection, cell-to-cell spread, and interactions with host cells (Leung et al., 2013). Since *inlD* is related to invasion activity (Seveau et al., 2007), serotype 1/2a and 1/2b strains are likely possess greater invasion ability than 1/2c strains. The distribution of internalins such as *inlD* and *inlG* suggests that formation of internalin clusters may be a multiple-step event and their acquisition could be related to divergence of these subgroups.

Genomic islands also play a vital role in survival and virulence of *L. monocytogenes* and their distribution appears to underpin different phenotypes observed in virulence among various subgroups. SSI-1, which was present in several subgroups (Ic, Id, IIe, IIg, III, IIj, and IIk), contributed to survival under low pH and high salt concentration (Ryan et al., 2010) and help *L. monocytogenes* to pass through the stomach and gut (Rychli et al., 2014). Moreover, LIPI-3 is likely to be acquired independently in subgroups Ia and Ic, known to encode hemolytic and cytotoxic factors as well as contributing to virulence of *L. monocytogenes* (Cotter et al., 2008). A total of 29 genomes including 12 from China contained *comK* phage junction fragment. It is noteworthy that the sequence variation in *comK* junction fragment may account for rapid adaptation and persistence in food processing plants and the contamination of RTE meats (Verghese et al., 2011).

The ncRNAs distribution also highlighted potential differences in *L. monocytogenes* pathogenesis (Mraheil et al., 2011). *rli29*, *rli85*, and *rliC* were present in EGDe (III) and 1/2c strains (IIj), all of which related to intracellular up-regulation in macrophage, although *rliC* was reported down-regulated in lineage III strains (Deng et al., 2010). These three ncRNAs are likely to be acquired by the common ancestor of subgroups III and IIj. Certain subgroups (IIC, IID, IIg, III, IIj, and IIk) carried *rli50*, documented previously as being involved in virulence in mice cells (Toledo-Arana et al., 2009; Mraheil et al., 2011).

In addition, strains in the same subgroups with small core SNP differences had different phenotypes in virulence due to possibly some strain specific genes (Table S4). SHL007 contained genes

encoding hemolysins (locus tags: I615_15111 and I615_15116) compared to SHL012; SHL013 carried multidrug efflux protein (I622_01390), ABC transporter proteins (I622_03330 and I622_03805) and flagella proteins (I622_05059 and I622_05069) compared to FSL J2-003.

The absence of *cas* gene clusters was not uncommon in *L. monocytogenes* (Hain et al., 2012; Kuenne et al., 2013). This lack of enzyme encoding DNA in these strains would likely render the entire CRISPR system unfunctional, which could facilitate the strains in these subgroups in acquiring foreign genetic elements including antibiotic resistance genes (Palmer and Gilmore, 2010). As the difficulty of assembling *cas* genes region in our draft genome, the findings here may be inconsistent.

However, as we got draft genomes for all strains, the limitation for the current study includes sequence gap, sequence error, and could achieve better genome quality with alternative assembly methods.

In summary, *L. monocytogenes* strains from China displayed a divergent population structure. Links between clinical and food strains, and a possible connection of strains from China and those associated with outbreak in another country indicate that whole genome sequencing data provide valuable information in public health and basic research. The differences in the distribution of virulence factors among *L. monocytogenes* from various sources highlighted variations in pathogenicity and the importance of horizontal gene transfer in the evolution and divergence of *L. monocytogenes*. The great resolution and power of whole genome sequencing advances a better and deeper understanding of the origins, emergence, and relationship of the dangerous foodborne pathogens.

DATA DEPOSITION

This project has been deposited at the Nucleotide database at NCBI under the GenBank accession numbers listed in **Table 2**.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: JM, JZ, and GC. Performed the experiments: JZ, GC, XY, and HP. Contributed

reagents/materials/analysis tools: XX, MA, EB, PL. Wrote the paper: JZ, GC.

FUNDING

This study was supported in part by the National Science and Technology Key Project (No. 2012ZX10004215-003), China-U.S. Collaborative Program on Emerging and Re-emerging Infectious Diseases (1U2GGH000961-01&5U2GGH000961-02) and the Joint Institute for Food Safety & Applied Nutrition, University of Maryland.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2016.01138>

Figure S1 | Genetic organization of SHL004, SHL015, 08-5578, and 08-5923. Strain 08-5578 was selected as reference sequence in the inner circle.

Figure S2 | Sequence comparison of plasmids pLMR479a, pLM5578, and those from strains SHL004 and SHL015.

Table S1 | Pairwise distance matrix of SNP number differences with stand deviation for 11 subgroups of compared genomes.

Table S2 | Pairwise distance matrix of SNP number differences with stand deviation for 41 compared genomes Strain.

Table S3 | The distribution of 117 genes related to virulence, regulation, and metabolism.

Table S4 | Strain specific genes identified from the comparison between strains with close relationship within the same subgroups. Strain specific genes in SHL007 compared to SHL012 (subgroup 1b).

Table S5 | Distribution of ncRNAs in all compared genomes.

Table S6 | Distribution of secrete systems proteins in all compared genomes.

Table S7 | General characterizations of phage B025 using complete genome C1-387 as example.

Table S8 | General characterizations of LGI-2 using complete genome SLCC2755 as example.

Table S9 | CRISPR arrays sequence in compared genomes.

REFERENCES

- Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2
- Becavin, C., Bouchier, C., Lechat, P., Archambaud, C., Creno, S., Gouin, E., et al. (2014). Comparison of widely used *Listeria monocytogenes* strains EGD, 10403S, and EGD-e highlights genomic variations underlying differences in pathogenicity. *MBio* 5, e00969–e00914. doi: 10.1128/mBio.00969-14
- Bergmann, S., Beard, P. M., Pasche, B., Lienenklaus, S., Weiss, S., Gahan, C. G., et al. (2013). Influence of internalin A murinisation on host resistance to orally acquired listeriosis in mice. *BMC Microbiol.* 13:90. doi: 10.1186/1471-20-13-90
- Bierne, H., Sabet, C., Personnic, N., and Cossart, P. (2007). Internalins: a complex family of leucine-rich repeat-containing proteins in *Listeria monocytogenes*. *Microbes Infect.* 9, 1156–1166. doi: 10.1016/j.micinf.2007.05.003
- Cantinelli, T., Chenal-Francisque, V., Diancourt, L., Frezal, L., Leclercq, A., Wirth, T., et al. (2013). Epidemic clones of *Listeria monocytogenes* are widespread and ancient clonal groups. *J. Clin. Microbiol.* 51, 3770–3779. doi: 10.1128/JCM.01874-13
- Cao, G., Meng, J., Strain, E., Stones, R., Pettengill, J., Zhao, S., et al. (2013). Phylogenetics and differentiation of *Salmonella* Newport lineages by whole genome sequencing. *PLoS ONE* 8:e55687. doi: 10.1371/journal.pone.0055687
- Clayton, E. M., Daly, K. M., Guinane, C. M., Hill, C., Cotter, P. D., and Ross, P. R. (2014). Atypical *Listeria innocua* strains possess an intact LIPI-3. *BMC Microbiol.* 14:58. doi: 10.1186/1471-2180-14-58
- Cotter, P. D., Draper, L. A., Lawton, E. M., Daly, K. M., Groeger, D. S., Casey, P. G., et al. (2008). Listeriolysin S, a novel peptide haemolysin associated with

- a subset of lineage I *Listeria monocytogenes*. *PLoS Pathog* 4:e1000144. doi: 10.1371/journal.ppat.1000144
- Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5:e11147. doi: 10.1371/journal.pone.0011147
- den Bakker, H. C., Desjardins, C. A., Griggs, A. D., Peters, J. E., Zeng, Q., Young, S. K., et al. (2013). Evolutionary dynamics of the accessory genome of *Listeria monocytogenes*. *PLoS ONE* 8:e67511. doi: 10.1371/journal.pone.0067511
- Deng, X., Phillippy, A. M., Li, Z., Salzberg, S. L., and Zhang, W. (2010). Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. *BMC Genomics* 11:500. doi: 10.1186/1471-2164-11-500
- Desvaux, M., and Hebraud, M. (2006). The protein secretion systems in *Listeria*: inside out bacterial virulence. *FEMS Microbiol. Rev.* 30, 774–805. doi: 10.1111/j.1574-6976.2006.00035.x
- Dussurget, O., Pizarro-Cerda, J., and Cossart, P. (2004). Molecular determinants of *Listeria monocytogenes* virulence. *Annu. Rev. Microbiol.* 58, 587–610. doi: 10.1146/annurev.micro.57.030502.090934
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461
- Gilmour, M. W., Graham, M., Van Domselaar, G., Tyler, S., Kent, H., Trout-Yakel, K. M., et al. (2010). High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. *BMC Genomics* 11:120. doi: 10.1186/1471-2164-11-120
- Gonzalez-Zorn, B., Dominguez-Bernal, G., Suarez, M., Ripio, M. T., Vega, Y., Novella, S., et al. (2000). SmcL, a novel membrane-damaging virulence factor in *Listeria*. *Int. J. Med. Microbiol.* 290, 369–374. doi: 10.1016/S1438-4221(00)80044-2
- Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* 35, W52–W57. doi: 10.1093/nar/gkm360
- Hain, T., Ghai, R., Billion, A., Kuenne, C. T., Steinweg, C., Izar, B., et al. (2012). Comparative genomics and transcriptomics of lineages I, II, and III strains of *Listeria monocytogenes*. *BMC Genomics* 13:144. doi: 10.1186/1471-2164-13-144
- Hein, I., Klinger, S., Dooms, M., Flekna, G., Stessl, B., Leclercq, A., et al. (2011). Stress survival islet 1 (SSI-1) survey in *Listeria monocytogenes* reveals an insert common to *Listeria innocua* in sequence type 121 L. monocytogenes strains. *Appl. Environ. Microbiol.* 77, 2169–2173. doi: 10.1128/AEM.02159-10
- Izar, B., Mraheil, M. A., and Hain, T. (2011). Identification and role of regulatory non-coding RNAs in *Listeria monocytogenes*. *Int. J. Mol. Sci.* 12, 5070–5079. doi: 10.3390/ijms12085070
- Jackson, K. A., Biggerstaff, M., Tobin-D'Angelo, M., Sweat, D., Klos, R., Nosari, J., et al. (2011). Multistate outbreak of *Listeria monocytogenes* associated with Mexican-style cheese made from pasteurized milk among pregnant, Hispanic women. *J. Food Prot.* 74, 949–953. doi: 10.4315/0362-028X.JFP-10-536
- Ju, W., Rump, L., Toro, M., Shen, J., Cao, G., Zhao, S., et al. (2014). Pathogenicity islands in Shiga toxin-producing *Escherichia coli* O26, O103, and O111 isolates from humans and animals. *Foodborne Pathog. Dis.* 11, 342–345. doi: 10.1089/fpd.2013.1696
- Klimke, W., Agarwala, R., Badretdin, A., Chetvernin, S., Ciufo, S., Fedorov, B., et al. (2009). The national center for biotechnology information's protein clusters database. *Nucleic Acids Res.* 37, D216–D223. doi: 10.1093/nar/gkn734
- Kuenne, C., Billion, A., Mraheil, M. A., Strittmatter, A., Daniel, R., Goesmann, A., et al. (2013). Reassessment of the *Listeria monocytogenes* pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome. *BMC Genomics* 14:47. doi: 10.1186/1471-2164-14-47
- Laksanalamai, P., Joseph, L. A., Silk, B. J., Burall, L. S., C, L.T., Gerner-Smidt, P., et al. (2012). Genomic characterization of *Listeria monocytogenes* strains involved in a multistate listeriosis outbreak associated with cantaloupe in US. *PLoS ONE* 7:e42448. doi: 10.1371/journal.pone.0042448
- Leung, N., Gianfelice, A., Gray-Owen, S. D., and Ireton, K. (2013). Impact of the *Listeria monocytogenes* protein InIC on infection in mice. *Infect. Immun.* 81, 1334–1340. doi: 10.1128/IAI.01377-12
- Mraheil, M. A., Billion, A., Mohamed, W., Mukherjee, K., Kuenne, C., Pischimarov, J., et al. (2011). The intracellular sRNA transcriptome of *Listeria monocytogenes* during growth in macrophages. *Nucleic Acids Res.* 39, 4235–4248. doi: 10.1093/nar/gkr033
- Neves, D., Job, V., Dortet, L., Cossart, P., and Dessen, A. (2013). Structure of internalin InlK from the human pathogen *Listeria monocytogenes*. *J. Mol. Biol.* 425, 4520–4529. doi: 10.1016/j.jmb.2013.08.010
- Palmer, K. L., and Gilmore, M. S. (2010). Multidrug-resistant enterococci lack CRISPR-cas. *MBio* 1:e00227-10. doi: 10.1128/mBio.00227-10
- Ragon, M., Wirth, T., Holland, F., Lavenir, R., Lecuit, M., Le Monnier, A., et al. (2008). A new perspective on *Listeria monocytogenes* evolution. *PLoS Pathog.* 4:e1000146. doi: 10.1371/journal.ppat.1000146
- Ryan, S., Begley, M., Hill, C., and Gahan, C. G. (2010). A five-gene stress survival islet (SSI-1) that contributes to the growth of *Listeria monocytogenes* in suboptimal conditions. *J. Appl. Microbiol.* 109, 984–995. doi: 10.1111/j.1365-2672.2010.04726.x
- Rychli, K., Muller, A., Zaiser, A., Schoder, D., Allerberger, F., Wagner, M., et al. (2014). Genome sequencing of *Listeria monocytogenes* Quargel listeriosis outbreak strains reveals two different strains with distinct *in vitro* virulence potential. *PLoS ONE* 9:e89964. doi: 10.1371/journal.pone.0089964
- Sabet, C., Lecuit, M., Cabanes, D., Cossart, P., and Bierne, H. (2005). LPXTG protein InlJ, a newly identified internalin involved in *Listeria monocytogenes* virulence. *Infect. Immun.* 73, 6912–6922. doi: 10.1128/IAI.73.10.6912-6922.2005
- Schoder, D., Stessl, B., Szakmary-Brandle, K., Rossmanith, P., and Wagner, M. (2014). Population diversity of *Listeria monocytogenes* in quargel (acid curd cheese) lots recalled during the multinational listeriosis outbreak 2009/2010. *Food Microbiol.* 39, 68–73. doi: 10.1016/j.fm.2013.11.006
- Seveau, S., Pizarro-Cerda, J., and Cossart, P. (2007). Molecular mechanisms exploited by *Listeria monocytogenes* during host cell invasion. *Microbes Infect.* 9, 1167–1175. doi: 10.1016/j.micinf.2007.05.004
- Smith, B., Larsson, J. T., Lisby, M., Muller, L., Madsen, S. B., Engberg, J., et al. (2011). Outbreak of listeriosis caused by infected beef meat from a meals-on-wheels delivery in Denmark 2009. *Clin. Microbiol. Infect.* 17, 50–52. doi: 10.1111/j.1469-0961.2010.03200.x
- Swaminathan, B., and Gerner-Smidt, P. (2007). The epidemiology of human listeriosis. *Microbes Infect.* 9, 1236–1243. doi: 10.1016/j.micinf.2007.05.011
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121
- Toledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., et al. (2009). The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature* 459, 950–956. doi: 10.1038/nature08080
- Touchon, M., and Rocha, E. P. (2010). The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS ONE* 5:e11126. doi: 10.1371/journal.pone.0011126
- Vergheese, B., Lok, M., Wen, J., Alessandria, V., Chen, Y., Kathariou, S., et al. (2011). comK prophage junction fragments as markers for *Listeria monocytogenes* genotypes unique to individual meat and poultry processing plants and a model for rapid niche-specific adaptation, biofilm formation, and persistence. *Appl. Environ. Microbiol.* 77, 3279–3292. doi: 10.1128/AEM.00546-11
- Wang, Y., Zhao, A., Zhu, R., Lan, R., Jin, D., Cui, Z., et al. (2012). Genetic diversity and molecular typing of *Listeria monocytogenes* in China. *BMC Microbiol.* 12:119. doi: 10.1186/1471-2180-12-119
- Zwickl, D. J. (2006). *Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion*. Austin, TX: The University of Texas at Austin.
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Copyright © 2016 Zhang, Cao, Xu, Allard, Li, Brown, Yang, Pan and Meng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genome Sequence of *Vibrio parahaemolyticus* VP152 Strain Isolated from *Penaeus indicus* in Malaysia

Vengadesh Letchumanan^{1,2}, Hooi-Leng Ser², Wen-Si Tan¹, Nurul-Syakima Ab Mutalib³, Bey-Hing Goh^{2,4}, Kok-Gan Chan^{1*} and Learn-Han Lee^{2,4*}

¹ Division of Genetics and Molecular Biology, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia, ² Novel Bacteria and Drug Discovery Research Group, School of Pharmacy, Monash University Malaysia, Bandar Sunway, Malaysia, ³ UKM Medical Molecular Biology Institute, UKM Medical Centre, Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia, ⁴ Center of Health Outcomes Research and Therapeutic Safety, School of Pharmaceutical Sciences, University of Phayao, Phayao, Thailand

Keywords: *Vibrio parahaemolyticus*, seafood, *Penaeus indicus*, antibiotic resistance, genome

OPEN ACCESS

Edited by:

Jennifer Ronholm,
Health Canada, Canada

Reviewed by:

Hongxia Wang,
University of Alabama at Birmingham,
USA

Jessica L. Jones,
United States Food and Drug
Administration, USA

*Correspondence:

Learn-Han Lee
lee.learn.han@monash.edu
leelearnhan@yahoo.com
Kok-Gan Chan
kokgan@um.edu.my

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 19 July 2016

Accepted: 25 August 2016

Published: 07 September 2016

Citation:

Letchumanan V, Ser H-L, Tan W-S, Ab Mutalib N-S, Goh B-H, Chan K-G and Lee L-H (2016) Genome Sequence of *Vibrio parahaemolyticus* VP152 Strain Isolated from *Penaeus indicus* in Malaysia. *Front. Microbiol.* 7:1410. doi: 10.3389/fmicb.2016.01410

INTRODUCTION

Vibrio parahaemolyticus is a Gram-negative bacterium that naturally occurs in marine associated aquatic environments (Letchumanan et al., 2014; Malcolm et al., 2015). This bacterium causes highest number of seafood-associated gastroenteritis in many countries including United States and Asian countries (Scallan et al., 2011; Newton et al., 2012). *V. parahaemolyticus* is often been isolated from aquatic environments such as seawater and marine sediment, as well as from vertebrate and invertebrate seafood products (Shen et al., 2009). The most likely route of infection in humans is reported to be associated with consumption of raw or improperly cooked seafood (Daniels et al., 2000; Jun et al., 2014; Hazen et al., 2015; Raghunath, 2015; Law et al., 2015).

Recently, *V. parahaemolyticus* has been demonstrated to be a major source of infection in the aquaculture industry (Letchumanan et al., 2014; Soto-Rodriguez et al., 2015; Tey et al., 2015). Aquaculture farmers rely on a wide range of antibiotics to prevent (prophylactic use) and treat (therapeutic use) bacterial infections in fish and invertebrates (Cabello et al., 2013). The extensive use of antibiotics and other chemotherapeutics in aquaculture has led to the emergence of multidrug resistant strains in the biosphere (Letchumanan et al., 2015a, 2016; Rao and Lalitha, 2015). Multidrug resistant *V. parahaemolyticus* strains have been isolated and detected from shrimp in Thailand (Yano et al., 2014), Malaysia (Al-Othribi et al., 2011; Sani et al., 2013; Letchumanan et al., 2015b,c) and China (Peng et al., 2010; Xu et al., 2014). Resistance toward clinically used antibiotics will eventually hamper the treatment of bacterial infections in humans and potentially increase the fatality rate (Daniels et al., 2000). Therefore, monitoring *Vibrio* species in aquaculture surroundings is crucial for both human health and the aquaculture industry.

In our previous study, we have isolated environmental *V. parahaemolyticus* strains from two types of Malaysian shrimp, *Penaeus indicus* and *Solenocera subnuda*. We detected the thermostable direct hemolysin (*tdh*) and thermostable direct related hemolysin (*trh*) virulence genes through a PCR based assay and studied the antibiotic resistance profile of all the isolated strains (Letchumanan et al., 2015c). *V. parahaemolyticus* VP152 was isolated from *Penaeus indicus* (Banana prawn) and originated from a supermarket sample. This strain did not possess both the *tdh* and *trh* virulence genes, which are responsible for causing diseases in humans and marine animals. Despite the fact that *V. parahaemolyticus* VP152 strain does not have *tdh* and *trh* virulence genes properties, the strain cannot be ignored in light of the fact that it exhibits multidrug resistance profiles toward 11/14 antibiotics tested. Based on the antibiotic

susceptibility phenotype, the strain exhibited multiple-antibiotic resistance toward ampicillin, oxytetracycline, nalidixic acid, ampicillin/sulbactam, tetracycline, third generation cephalosporins (cefotaxime and ceftazidime), aminoglycosides (amikacin, kanamycin, and gentamicin) and trimethoprim/sulfamethoxazole (Letchumanan et al., 2015c).

This is a worrying situation as the antibiotic resistant profiles shown by *V. parahaemolyticus* VP152 include the recommended antimicrobial agents used in treatment of *Vibrio* spp. infections, including third generation cephalosporin, fluoroquinolones, aminoglycosides, tetracycline, gentamicin, trimethoprim/sulfamethoxazole (Daniels and Shafaie, 2000; Shaw et al., 2014). Therefore, the whole genome sequence of *V. parahaemolyticus* VP152 was studied with respect to the multidrug resistance profiles to gain a better understanding of the antibiotic resistant patterns. The availability of this genome sequence of *V. parahaemolyticus* VP152 will aid as a basis for further in-depth analysis of the antibiotic resistance profile of environmental *V. parahaemolyticus*.

MATERIALS AND METHODS

Genome Sequencing and Assembly

Genomic DNA of VP152 strain was extracted using MasterpureTM DNA purification kit (Epicenter, Illumina Inc, Madison, WI, USA) and subjected to RNase (Qiagen, USA) treatment (Ser et al., 2015). The DNA quality was quantified using NanoDrop spectrophotometer (Thermo Scientific, Waltham, MA, USA), and a Qubit version 2.0 fluorometer (Life Technologies, Carlsbad, CA, USA). Illumina sequencing library of genomic DNA was prepared using NexteraTM DNA Sample Preparation kit (Illumina, San Diego, CA, USA) and library quality was validated by a Bioanalyzer 2100 high sensitivity DNA kit (Agilent Technologies, Palo Alto, CA, USA) prior to sequencing. The genome of VP152 strain was sequenced on MiSeq platform with MiSeq Reagent Kit 2 (2 × 250 bp; Illumina Inc, San Diego, CA, USA). The trimmed sequences were *de novo* assembled with CLC Genomic Workbench version 5.1 (CLC Bio, Denmark).

Genome Annotation

Gene prediction was carried out using Prodigal 2.6, while rRNA and tRNA were analyzed using RNAmmer and tRNAscan SE version 1.21 (Lowe and Eddy, 1997; Lagesen et al., 2007; Hyatt et al., 2010). Gene prediction and annotation were performed using Rapid Annotation Search Tool (RAST; Aziz et al., 2008). Antibiotic resistance genes were analyzed using antibiotic resistance genes-ANNOTation (ARG-ANNOT; Gupta et al., 2014).

RESULTS

Genome Characteristics

The genome of *V. parahaemolyticus* VP152 consists of 4,982,021 bp with mean genome coverage of 183.46-fold

and with an average G+C content of 53.4% (Table 1). A total of 4809 genes was predicted of which 4638 were identified as protein coding genes. There are 91 RNA genes consisting of 11 rRNAs and 80 tRNAs.

Virulence and Antimicrobial Resistance Genes

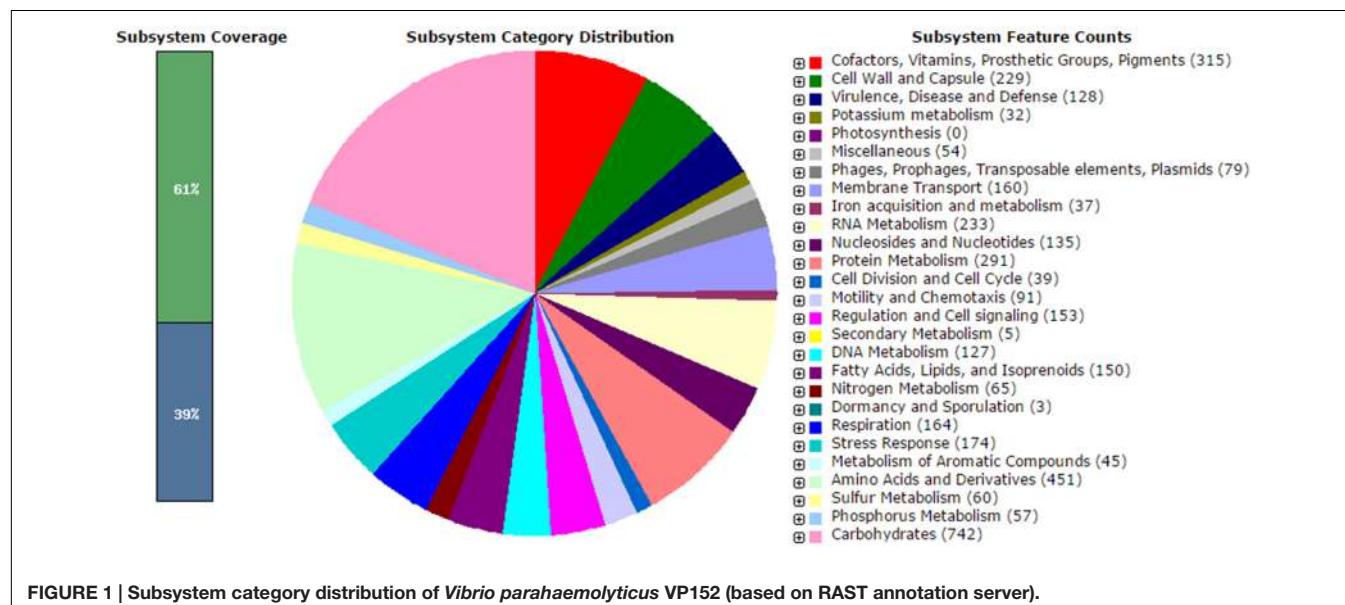
The analysis obtained from RAST server revealed 573 subsystems (Figure 1). The annotated genome has 97 genes responsible for resistance to antibiotic and toxic compounds including seven genes for mdtABCD multidrug resistance cluster, 19 genes for multidrug resistance efflux pumps, four genes for β-lactamase and two genes aminoglycoside adenyllyltransferases. The genome sequence of *V. parahaemolyticus* VP152 was compared with three environmental *V. parahaemolyticus* strains, in order to delineate the similarities between the four strains. The genome size of *V. parahaemolyticus* VP152 was similar to strains of *V. parahaemolyticus* and contained several antibiotic resistance genes as shown in Table 1. Also, further comparison of hemolysin genes present in *V. parahaemolyticus* VP152 and the selected strains revealed no significant differences.

The genome analysis on ARG-ANNOT noted the presences of tetracycline resistant gene, *Tet* and *Tet-2* gene within the genome. The presence of these genes is closely related to the phenotypic resistance shown by the strain toward oxytetracycline and tetracycline. Furthermore, β-lactam resistance-related gene, *bla* gene of VP152 exhibited 99% similarities when compared to other *V. parahaemolyticus* strain and *Vibrio* species. The phenotypic resistance shown by *V. parahaemolyticus* VP152 toward ampicillin, ampicillin/sulbactam, cefotaxime and ceftazidime is closely related to the gene coding β-lactamase in the genome. The gene coding aminoglycosides adenyllyltransferase of *V. parahaemolyticus* VP152 confers resistance phenotype observed toward amikacin, kanamycin, and gentamicin. Based on the annotation tools and detailed analysis of *V. parahaemolyticus* VP152 genome using PlasmidFinder, the genome of *V. parahaemolyticus* VP152 did not recover any plasmid sequence. Even though these genes were commonly found in plasmids, some of the *Vibrio* species including *V. corallilyticus* and *V. alginolyticus* carry these genes in their chromosomes (Costa et al., 2015). Therefore, the resistant genes observed in *V. parahaemolyticus* VP152 are chromosome mediated.

The multidrug resistance profile seen in the phenotype and genes of *V. parahaemolyticus* VP152 genome illustrates how extensive antibiotics have been utilized in the aquaculture industry. The resistance phenotype observed in this strain could be triggered by the extensive use of permitted antibiotics in the Asian aquaculture industry namely oxytetracycline, tetracycline, quinolone, sulphonamides, and trimethoprim (Rico et al., 2012; Yano et al., 2014). The resistance toward third generation cephalosporins seen in *V. parahaemolyticus* VP152 would further hamper the treatment of *Vibrio* species infection in future. This situation is cause for concern, and warrants more stringent surveillance in the use of antibiotics, as well

TABLE 1 | Comparison of genome sequence of *Vibrio parahaemolyticus* VP152 with other genome sequences.

	<i>Vibrio parahaemolyticus</i> VP152	<i>Vibrio parahaemolyticus</i> VP551	<i>Vibrio parahaemolyticus</i> M0605	<i>Vibrio parahaemolyticus</i> AQ4037
Source of isolation	Shrimp	Water source	Environmental	Shrimp
Genome size (bp)	4,982,021	5,226,872	5,429,407	4,939,804
Genome coverage (fold)	183.46	256.00	20.00	7.37
Contig N ₅₀ (bp)	566,732	712,378	121,988	67,710
Sequencing technology	Illumina MiSeq	SOLiD	Ion Torrent	Sanger
KEGG categories, number of genes (genome %)	61 (1.91)	49 (1.73)	46 (1.71)	49 (1.71)
Cationic antimicrobial peptide (CAMP) resistance, number of genes	36	21	23	20
Vancomycin resistance, number of genes	8	7	7	7
β-Lactam resistance, number of genes	20	27	22	28

**FIGURE 1 | Subsystem category distribution of *Vibrio parahaemolyticus* VP152 (based on RAST annotation server).**

as the resultant antibiotic resistance in clinically important bacterial species. In summary, the whole genome sequence of *V. parahaemolyticus* VP152 will be useful in future studies to determine antimicrobial resistance and virulence attributes as well as mechanisms that enhance its environmental or host fitness.

Nucleotide Sequence Accession Numbers

This genome sequence data of VP152 strain sequenced under this study has been deposited in DDBJ/EMBL/GenBank under Accession No. LCUL00000000. The version described in this paper is the first version, LCUL01000000. The genome sequences data are available in FASTA, annotated GenBank flat file, graphical and ASN.1 formats.

AUTHOR CONTRIBUTIONS

The experiments, data analysis and manuscript writing were performed by VL and H-LS, while W-ST, N-SA, B-HG, K-GC, and L-HL provided vital guidance, technical support, and proofreading for the work. The research project was founded by L-HL.

ACKNOWLEDGMENTS

This work was supported by University of Malaya for High Impact Research Grant (UM-MOHE HIR Nature Microbiome Grant No. H-50001-A000027) awarded to K-GC., MOSTI eScience Fund (06-02-10-SF0300) and External Industry Grants from Bitek Abadi Sdn Bhd (vote no. GBA-808138 & GBA-808813) awarded to L-HL.

REFERENCES

- Al-Othrubi, S. M., Alfizah, H., Son, R., Humin, N., and Rahaman, J. (2011). Rapid detection and E-test antimicrobial susceptibility testing of *Vibrio parahaemolyticus* isolated from seafood and environmental sources in Malaysia. *Saudi Medic. J.* 32, 400–406.
- Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75
- Cabello, F. C., Godfrey, H. P., Tomova, A., Ivanova, L., Dölz, H., Millanao, A., et al. (2013). Antimicrobial use in aquaculture re-examined: its relevance to antimicrobial resistance and to animal and human health. *Appl. Environ. Microbiol.* 15, 1917–1942. doi: 10.1111/1462-2920.12134
- Costa, R. A., Araujo, R. L., Souza, O. V., and Vieira, H. S. F. (2015). Antibiotic-resistant Vibrios in farmed shrimp. *BioMed. Res. Int.* 2015, 1–5. doi: 10.1155/2015/602078
- Daniels, N., and Shafaei, A. (2000). A review of pathogenic *Vibrio* infections for clinicians. *J. Infect. Medic.* 17, 665–685.
- Daniels, N. A., MacKinnon, L., Bishop, R., Altekkruse, S., Ray, B., Hammond, R. M., et al. (2000). *Vibrio parahaemolyticus* infections in the United States, 1973–1998. *J. Infect. Dis.* 181, 1661–1666. doi: 10.1086/315459
- Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Rojas, R. L., Kemf, M., Landraud, L., et al. (2014). ARG-ANNOT, a new bioinformatics tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemo.* 58, 212–220. doi: 10.1128/AAC.01310-13
- Hazen, T. H., Lafon, P. C., Garrett, N. M., Lowe, T. M., Silberger, D. J., Rowe, L. A., et al. (2015). Insights into the environmental reservoir of pathogenic *Vibrio parahaemolyticus* using comparative genomics. *Front. Microbiol.* 6:204. doi: 10.3389/fmicb.2015.00204
- Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119
- Jun, J. W., Kim, H. J., Yun, S. K., Choi, J. Y., and Park, S. C. (2014). Eating oysters without risk of vibriosis: application of a bacteriophage against *Vibrio parahaemolyticus* in oysters. *Int. J. Food Microbiol.* 188, 31–35. doi: 10.1016/j.ijfoodmicro.2014.07.007
- Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T., and Ussery, D. W. (2007). RNAMmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108. doi: 10.1093/nar/gkm160
- Law, J. W.-F., Ab Mutalib, N. S., Chan, K.-G., and Lee, L.-H. (2015). Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. *Front. Microbiol.* 5:770. doi: 10.3389/fmicb.2014.00770
- Letchumanan, V., Chan, K.-G., and Lee, L.-H. (2014). *Vibrio parahaemolyticus*: a review on the pathogenesis, prevalence and advance molecular identification techniques. *Front. Microbiol.* 5:705. doi: 10.3389/fmicb.2014.00705
- Letchumanan, V., Chan, K.-G., and Lee, L.-H. (2015a). An insight of traditional plasmid curing in *Vibrio* species. *Front. Microbiol.* 6:735. doi: 10.3389/fmicb.2015.00735
- Letchumanan, V., Chan, K.-G., Pusparajah, P., Saokaew, S., Duangjai, A., Goh, B.-H., et al. (2016). Insights into bacteriophage application in controlling *Vibrio* species. *Front. Microbiol.* 7:1114. doi: 10.3389/fmicb.2016.01114
- Letchumanan, V., Pusparajah, P., Loh, T. H. T., Yin, W.-F., Lee, L.-H., and Chan, K.-G. (2015b). Occurrence and antibiotic resistance of *Vibrio parahaemolyticus* from shellfish in Selangor, Malaysia. *Front. Microbiol.* 6:1417. doi: 10.3389/fmicb.2015.01417
- Letchumanan, V., Yin, W.-F., Lee, L.-H., and Chan, K.-G. (2015c). Prevalence and antimicrobial susceptibility of *Vibrio parahaemolyticus* isolated from retail shrimps in Malaysia. *Front. Microbiol.* 6:33. doi: 10.3389/fmicb.2015.00033
- Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.0955
- Malcolm, T. T. H., Cheah, Y. K., Radzi, C. W. J. W. M., Kasim, F. A., Kantilal, H. K., John, T. Y. H., et al. (2015). Detection and quantification of pathogenic *Vibrio parahaemolyticus* in shellfish by using multiplex PCR and loop-mediated isothermal amplification assay. *Food Cont.* 47, 664–671. doi: 10.1016/j.foodcont.2014.08.010
- Newton, A., Kendall, M., Vugia, D. J., Henao, O. L., and Mahon, B. E. (2012). Increasing rates of vibriosis in the United States, 1996–2010: review of surveillance data from 2 systems. *Clin. Infect. Dis.* 54, 391–395. doi: 10.1093/cid/cis243
- Peng, F. M., Jiang, D. Y., Ruan, H. H., Liu, H. Q., and Zhou, L. P. (2010). Pathogenic investigation on a food poisoning induced by *Vibrio parahaemolyticus*. *Prev. Med. Trib.* 16, 746–747.
- Raghunath, P. (2015). Roles of thermostable direct hemolysin (TDH) and TDH-related hemolysin (TRH) in *Vibrio parahaemolyticus*. *Front. Microbiol.* 5:805. doi: 10.3389/fmicb.2014.00805
- Rao, B. M., and Lalitha, K. V. (2015). Bacteriophage for aquaculture: are they beneficial or inimical. *Aquaculture* 437, 146–154. doi: 10.1016/j.aquaculture.2014.11.039
- Rico, A., Satapornvanit, K., Haque, M. M., Min, J., Nguyen, P. T., Telfer, T., et al. (2012). Use of chemicals and biological products in Asian aquaculture and their potential environmental risks: a critical review. *Rev. Aqua.* 4, 75–93. doi: 10.1111/j.1753-5131.2012.01062.x
- Sani, N. A., Ariyawansa, S., Babji, A. S., and Hashim, J. K. (2013). The risk assessment of *Vibrio parahaemolyticus* in cooked black tiger shrimps (*Penaeus monodon*) in Malaysia. *Food Cont.* 31, 546–552. doi: 10.1016/j.foodcont.2012.10.018
- Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., and Roy, S. L. (2011). Foodborne illness acquired in the United States e major pathogens. *Emerg. Infect. Dis.* 17, 7–15. doi: 10.3201/eid1707.110572
- Ser, H. L., Tan, W. S., Cheng, H. J., Yin, W. F., Chan, K. G., and Lee, L. H. (2015). Draft genome of amylolytic actinobacterium, *Sinomonas humi* MUSC 117T isolated from intertidal soil. *Mar. Genomics* 24, 209–210. doi: 10.1016/j.margen.2015.05.012
- Shaw, K. S., Goldstein, R. E. R., He, X., Jacobs, J. M., Crump, B. C., and Sapkota, A. R. (2014). Antimicrobial susceptibility of *Vibrio vulnificus* and *Vibrio parahaemolyticus* recovered from recreational and commercial areas of Chesapeake Bay and Maryland Coastal Bays. *PLoS ONE* 9:89616. doi: 10.1371/journal.pone.0089616
- Shen, X., Cai, Y., Liu, C., Liu, W., Hui, Y., and Su, Y.-C. (2009). Effect of temperature on uptake and survival of *Vibrio parahaemolyticus* in oysters (*Crassostrea plicatula*). *Int. J. Food. Microbiol.* 136, 129–132. doi: 10.1016/j.ijfoodmicro.2009.09.012
- Soto-Rodriguez, S. A., Gomez-Gil, B., Lozano-Olvera, R., Betancourt-Lozano, M., and Morales-Covarrubias, M. S. (2015). Field and experimental evidence of *Vibrio parahaemolyticus* as the causative agent of acute hepatopancreatic necrosis disease of cultured shrimp (*Litopenaeus vannamei*) in northwestern Mexico. *Appl. Environ. Microbiol.* 81, 1689–1699. doi: 10.1128/AEM.03610-14
- Tey, Y. H., Jong, K. J., Fen, S. Y., and Wong, H. C. (2015). Occurrence of *Vibrio parahaemolyticus*, *Vibrio cholerae*, and *Vibrio vulnificus* in the aquacultural environments of Taiwan. *J. Food Prot.* 78, 969–976. doi: 10.4315/0362-028X.JFP-14-405
- Xu, X., Wu, Q., Zhang, J., Cheng, J., Zhang, S., and Wu, K. (2014). Prevalence, pathogenicity, and serotypes of *Vibrio parahaemolyticus* in shrimp from Chinese retail markets. *Food Cont.* 46, 81–85. doi: 10.1016/j.foodcont.2014.04.042
- Yano, Y., Hamano, K., Satomi, M., Tsutsui, I., Ban, M., and Aue-umneoy, D. (2014). Prevalence and antimicrobial susceptibility of *Vibrio* species related to food safety isolated from shrimp cultured at inland ponds in Thailand. *Food Cont.* 38, 30–45. doi: 10.1016/j.fm.2014.11.003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Letchumanan, Ser, Tan, Ab Mutalib, Goh, Chan and Lee. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

Edited by:

Sabah Bidawid,
Health Canada, Canada

Reviewed by:

Young Min Kwon,
University of Arkansas, United States
Sheng Chen,
Hong Kong Polytechnic University,
Hong Kong

***Correspondence:**

Roger C. Levesque
rclevesq@ibis.ulaval.ca
Lawrence Goodridge
lawrence.goodridge@mcgill.ca

[†]These authors have contributed
equally to this work.

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 29 March 2017

Accepted: 17 May 2017

Published: 02 June 2017

Citation:

Emond-Rheault J-G, Jeukens J, Freschi L, Kukavica-Ibrulj I, Boyle B, Dupont M-J, Colavecchio A, Barrere V, Cadieux B, Arya G, Bekal S, Berry C, Burnett E, Cavestri C, Chapin TK, Crouse A, Daigle F, Danyluk MD, Delaquis P, Dewar K, Doualla-Bell F, Fliss I, Fong K, Fournier E, Franz E, Garduno R, Gill A, Gruenheid S, Harris L, Huang CB, Huang H, Johnson R, Joly Y, Kerhoas M, Kong N, Lapointe G, Larivière L, Loignon S, Malo D, Moineau S, Mottawea W, Mukhopadhyay K, Nadon C, Nash J, Ngueng Feze I, Ogunremi D, Perets A, Pilar AV, Reimer AR, Robertson J, Rohde J, Sanderson KE, Song L, Stephan R, Tambar S, Thomassin P, Tremblay D, Usongo V, Vincent C, Wang S, Weadge JT, Wiedmann M, Wijnands L, Wilson ED, Wittum T, Yoshida C, Youfsi K, Zhu L, Weimer BC, Goodridge L and Levesque RC (2017) A Syst-OMICS Approach to Ensuring Food Safety and Reducing the Economic Burden of Salmonellosis. *Front. Microbiol.* 8:996. doi: 10.3389/fmicb.2017.00996

A Syst-OMICS Approach to Ensuring Food Safety and Reducing the Economic Burden of Salmonellosis

Jean-Guillaume Emond-Rheault^{1†}, Julie Jeukens^{1†}, Luca Freschi^{1†}, Irena Kukavica-Ibrulj¹, Brian Boyle¹, Marie-Josée Dupont¹, Anna Colavecchio², Virginie Barrere², Brigitte Cadieux², Gitanjali Arya³, Sadja Bekal⁴, Chrystal Berry³, Elton Burnett², Camille Cavestri⁵, Travis K. Chapin⁶, Alanna Crouse², France Daigle⁷, Michelle D. Danyluk⁶, Pascal Delaquis⁸, Ken Dewar^{2,9}, Florence Doualla-Bell⁴, Ismail Fliss⁵, Karen Fong¹⁰, Eric Fournier⁴, Eelco Franz¹¹, Rafael Garduno¹², Alexander Gill¹³, Samantha Gruenheid², Linda Harris¹⁴, Carol B. Huang¹⁵, Hongsheng Huang¹⁶, Roger Johnson³, Yann Joly², Maud Kerhoas⁷, Nguyen Kong¹⁵, Gisèle Lapointe¹⁷, Line Larivière², Stéphanie Loignon⁵, Danielle Malo², Sylvain Moineau⁵, Walid Mottawea^{2,18}, Kakali Mukhopadhyay², Céline Nadon³, John Nash³, Ida Ngueng Feze², Dele Ogunremi¹⁶, Ann Perets³, Ana V. Pilar², Aleisha R. Reimer³, James Robertson³, John Rohde¹⁹, Kenneth E. Sanderson²⁰, Lingqiao Song², Roger Stephan²¹, Sandeep Tambar¹³, Paul Thomassin², Denise Tremblay⁵, Valentine Usongo⁴, Caroline Vincent⁴, Siyun Wang¹⁰, Joel T. Weadge²², Martin Wiedmann²³, Lucas Wijnands¹¹, Emily D. Wilson²², Thomas Wittum²⁴, Catherine Yoshida³, Khadija Youfsi⁴, Lei Zhu², Bart C. Weimer¹⁵, Lawrence Goodridge^{2*} and Roger C. Levesque^{1*}

¹ Institute for Integrative and Systems Biology, Université Laval, Québec City, QC, Canada, ² McGill University, Montréal, QC, Canada, ³ National Microbiology Laboratory, Public Health Agency of Canada, Ottawa, ON, Canada, ⁴ Laboratoire de Santé Publique du Québec, Sainte-Anne-de-Bellevue, QC, Canada, ⁵ Université Laval, Québec City, QC, Canada, ⁶ Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL, United States, ⁷ Département de Microbiologie, Infectiologie et Immunologie, Université de Montréal, Montréal, QC, Canada, ⁸ Agriculture and Agri-Food Canada, Summerland, BC, Canada, ⁹ Génome Québec Innovation Center, Montréal, QC, Canada, ¹⁰ Food Safety Engineering, Faculty of Land and Food Systems, University of British Columbia, Vancouver, BC, Canada, ¹¹ National Institute for Public Health and the Environment, Bilthoven, Netherlands, ¹² Canadian Food Inspection Agency, Halifax, NS, Canada, ¹³ Bureau of Microbial Hazards, Health Canada, Ottawa, ON, Canada, ¹⁴ UC Davis Food Science and Technology, Davis, CA, United States, ¹⁵ UC Davis School of Veterinary Medicine, Davis, CA, United States, ¹⁶ Canadian Food Inspection Agency, Ottawa, ON, Canada, ¹⁷ Food Science, University of Guelph, Guelph, ON, Canada, ¹⁸ Department of Microbiology and Immunology, Faculty of Pharmacy, Mansoura University, Mansoura, Egypt, ¹⁹ Department of Microbiology and Immunology, Dalhousie University, Halifax, NS, Canada, ²⁰ Department of Biological Sciences, University of Calgary, Calgary, AB, Canada, ²¹ Institute for Food Safety and Hygiene, University of Zurich, Zurich, Switzerland, ²² Biological and Chemical Sciences, Wilfrid Laurier University, Waterloo, ON, Canada, ²³ Department of Food Science, Cornell University, Ithaca, NY, United States, ²⁴ College of Veterinary Medicine, The Ohio State University, Columbus, OH, United States

The *Salmonella* Syst-OMICS consortium is sequencing 4,500 *Salmonella* genomes and building an analysis pipeline for the study of *Salmonella* genome evolution, antibiotic resistance and virulence genes. Metadata, including phenotypic as well as genomic data, for isolates of the collection are provided through the *Salmonella* Foodborne Syst-OMICS database (SalFoS), at <https://salfos.ibis.ulaval.ca/>. Here, we present our strategy and the analysis of the first 3,377 genomes. Our data will be used to draw potential links between strains found in fresh produce, humans, animals and the environment. The ultimate goals are to understand how *Salmonella* evolves over time, improve the accuracy of diagnostic methods, develop control methods in the field, and identify prognostic markers for evidence-based decisions in epidemiology and surveillance.

Keywords: *Salmonella*, foodborne pathogen, next-generation sequencing, bacterial genomics, phylogeny, antibiotic resistance, database

IMPORTANCE OF FOODBORNE *Salmonella* AS A MODEL IN LARGE-SCALE BACTERIAL GENOMICS

Salmonella enterica is a foodborne bacterial pathogen having at least 2,600 serotypes (Gal-Mor et al., 2014)¹ that contaminates a diversity of foods and is a leading cause of foodborne illnesses and mortality globally. In fact, there are an estimated 93.3 million cases of gastroenteritis due to non-typhoidal *Salmonella* infections each year, resulting in approximately 155,000 deaths (Majowicz et al., 2010). In Canada, non-typhoidal salmonellosis accounts for more than 88,000 cases of foodborne illness each year, and has among the highest incidence rate of any bacterial foodborne pathogen (Thomas et al., 2015). *S. enterica* is responsible for more than 50% of fresh produce-borne outbreaks, the highest number of foodborne outbreaks of any inspected food commodity in North America (Kozak et al., 2013). Because of its remarkable genomic diversity, *Salmonella* is found in complex environmental and ecological niches and survives in harsh environments for long periods (Podolak et al., 2010; Fatica and Schneider, 2011). Several research groups have identified relationships between some of the 2,557 *S. enterica* serotypes and specific foods, which suggests, that some food commodities act as reservoirs for particular serotypes (Kim, 2010; Jackson et al., 2013; Nuesch-Inderbinen et al., 2015).

Salmonella outbreaks are monitored with support from the PulseNet surveillance system in 86 countries² (Ribot and Hise, 2016; Scharff et al., 2016). PulseNet Canada³ is a national surveillance system used to quickly identify and respond to foodborne disease outbreaks, centralized at the National Microbiology Laboratory in Winnipeg, MB, and working in close collaboration with a network of federal and provincial public health laboratories and epidemiologists. Still, despite the availability of thousands of sequenced genomes, knowledge of genome evolution integrated with transmission and epidemiology is limited for produce-related outbreaks.

Studies of *S. enterica* population structure in humans, animals, food and the environment are central to understand the biodiversity, evolution, ecology and epidemiology of this pathogen. However, studies describing the genetic structure of *Salmonella* populations are commonly based on isolates drawn overwhelmingly from clinical collections (Hoffmann et al., 2014). This approach has resulted in a limited view of *Salmonella*'s evolutionary history (D'costa et al., 2006; Perry and Wright, 2014). In *Salmonella* as in many other bacterial pathogens, there is limited knowledge on how genome content, rearrangements and the complement of genes including those acquired by horizontal gene transfer (HGT) contribute to strain-specific phenotypes, including virulence (Casadevall, 2017). Various studies have sought to resolve the population structure of *Salmonella* using complementary subtyping methods including pulsed-field gel electrophoresis (PFGE), multiple

loci VNTR analysis (MLVA), 7-gene housekeeping schemes, whole-genome multi-locus sequence typing (wgMLST) profiles, pan- and core genome studies, and CRISPR analysis to define molecular signatures, pathogen subtypes and the potential for pathogenicity (Shariat and Dudley, 2014; Rouli et al., 2015; Liu et al., 2016). Next-generation sequencing (NGS) coupled with whole-genome comparison is well-positioned to become the gold standard subtyping method, as it offers previously unmatched resolution for phylogenetic analysis and rapid subtyping during investigation of food contamination and outbreaks (Ashton et al., 2016; Bekal et al., 2016).

THE Syst-OMICS Strategy

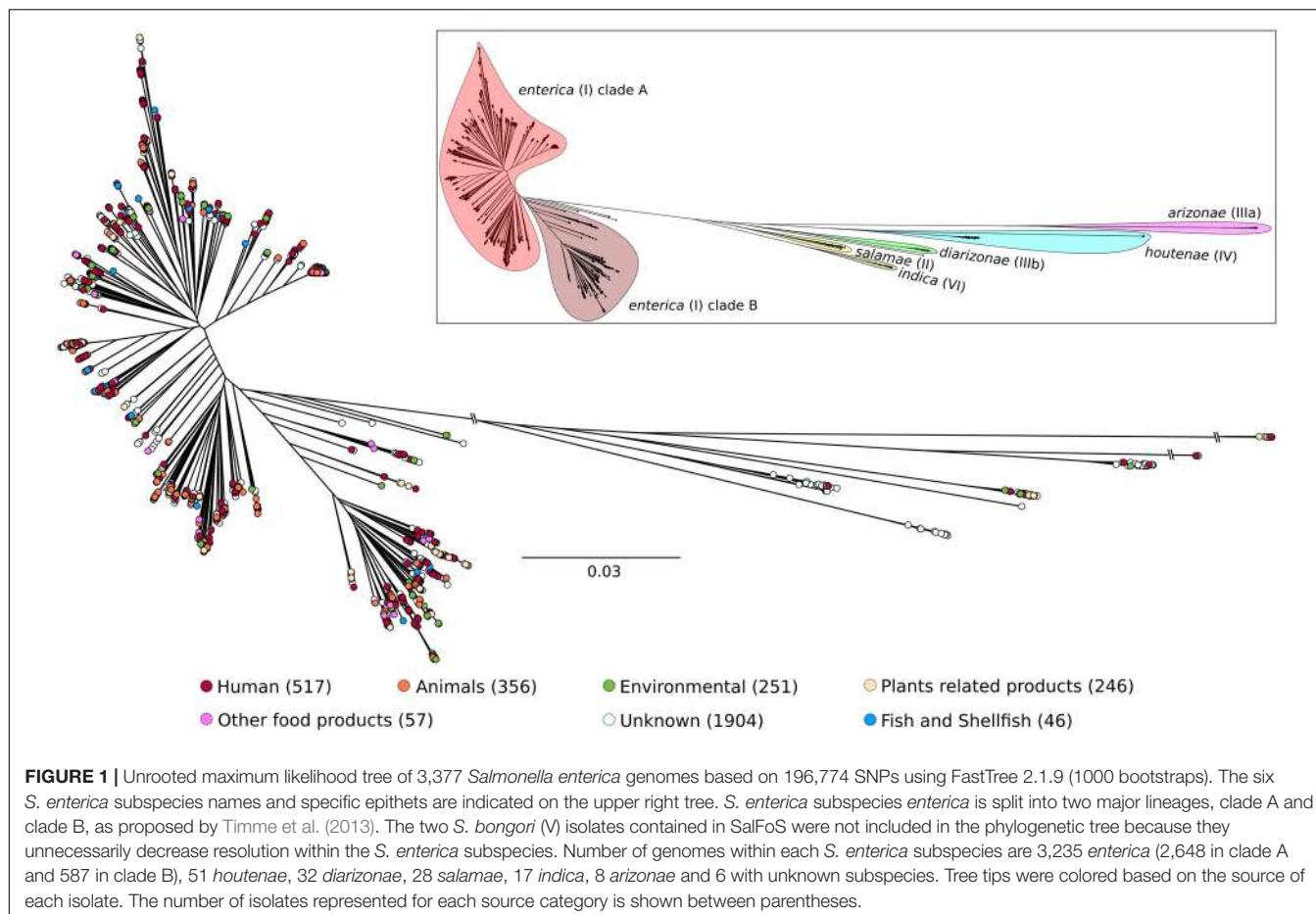
The application of genomics to infectious pathogens via WGS is transforming the practice of *Salmonella* diagnostics, epidemiology and surveillance. Genomic data are increasingly used to understand infectious disease epidemiology (Didelot et al., 2017). With rapidly falling costs and turnaround time, microbial WGS and analysis is becoming a viable strategy to identify the geographic origin of bacterial pathogens (Weedmark et al., 2015; Hoffmann et al., 2016). The objective of the Canadian-based international Syst-OMICS consortium is to sequence a minimum of 4,500 genomes, include the data in the *Salmonella* Foodborne Syst-OMICS database (SalFoS) at <https://SalFoS.ibis.ulaval.ca/>, share this information plus available metadata with Canadian federal and provincial regulators and the food industry, and develop pipelines to study these genomes. Genomics data will support molecular epidemiology and source attribution of outbreaks and has the potential for future genotypic antimicrobial susceptibility testing, as well as the identification of novel therapeutic targets and prognostic markers. Moreover, the large-scale genomics and evolutionary biology tools developed may lead to new strategies for countering not only *Salmonella* infections, but other pathogens as well (Little et al., 2012).

The Syst-OMICS project is based upon a systems approach (flowchart and screening method available in Supplementary File 1). First, the genome diversity of 4,500 isolates will be assessed using high-quality WGS, assembly, annotation and phylogeny. This data will be used for *in silico* serotyping (Yoshida et al., 2016), as well as analysis of virulence (Chen et al., 2012), antibiotic resistance (Jia et al., 2016) and mobilome gene content (Lanza et al., 2014). Based on this genomic data, a funnel-type model will be applied such that 300 isolates will be selected for *in vitro* high-throughput screening (HTS) in cell lines to determine attachment, adhesion, invasion and replication of each isolate (protocol adapted to 96-well plates from Forest et al., 2007). From the results, isolates will be categorized as being of high, medium, or low virulence. A limited number of those isolates will then be selected for further screening *in vivo* using a mouse model (Roy et al., 2007) and *in vitro* using gastrointestinal fermenter models (Kheadr et al., 2010; Le Blay et al., 2012). These data will identify isolates to represent the different levels of virulence that will be used to develop novel diagnostic and control tools. We propose to enhance food safety and lower the economic burden of salmonellosis through a farm-to-table

¹<https://www.cdc.gov/salmonella/reportspubs/salmonella-atlas/serotype-snapshots.html>

²<http://www.pulsenetinternational.org/networks/usa/>

³<https://www.nml-lnm.gc.ca/Pulsenet/index-eng.htm>



systematic approach to control *Salmonella*, with a focus on new control methods in agricultural production, more specific diagnostics and improved bacterial subtyping methods to support investigation of foodborne outbreaks, as no single intervention is likely to produce meaningful and lasting effects.

THE *Salmonella* FOODBORNE Syst-OMICS DATABASE (SalFoS)

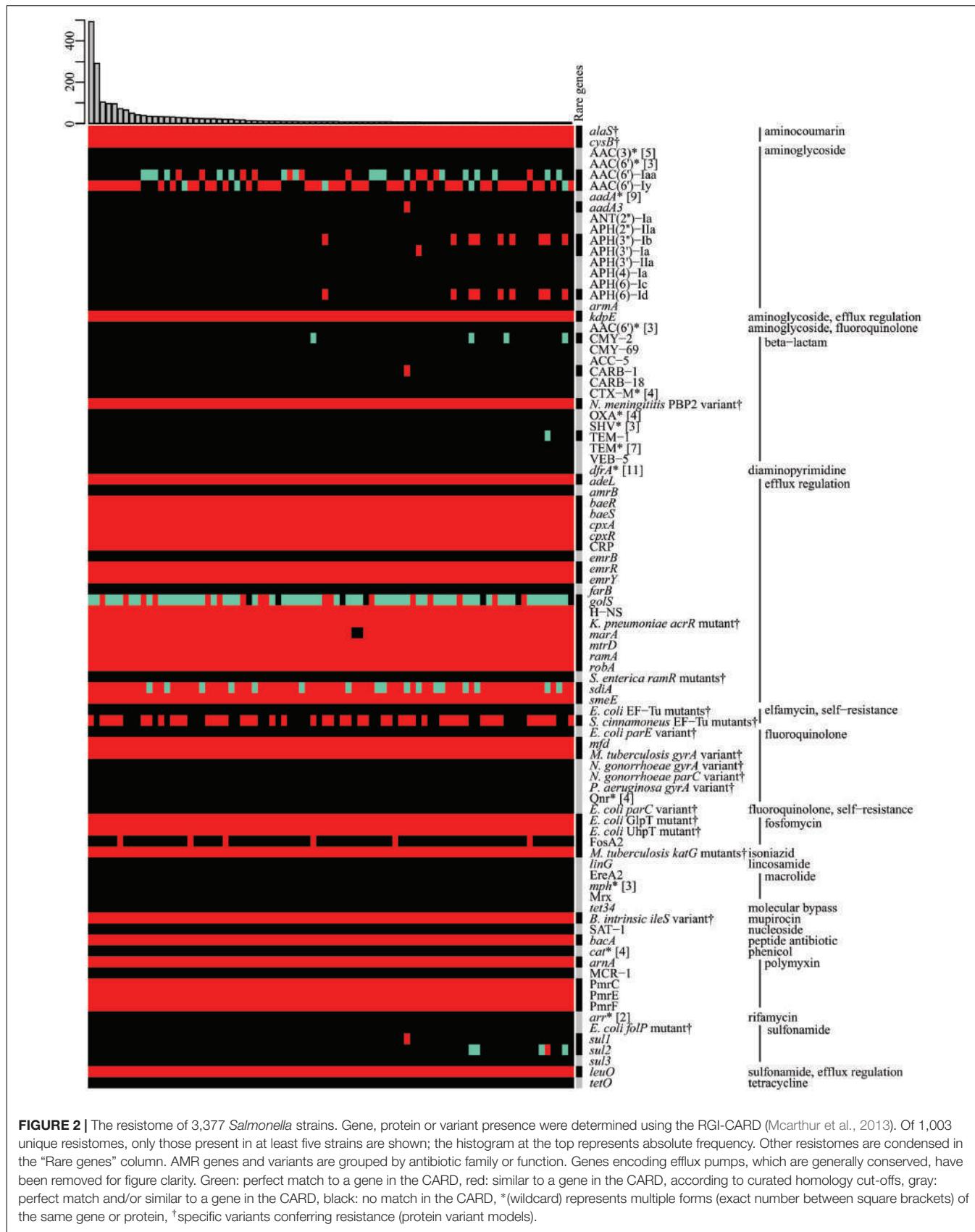
Salmonella Foodborne Syst-OMICS database is an online web application that relies on a Mysql 5 database. It was designed not only to store data for the *Salmonella* strain collection but also to provide access to each isolate's phenotypic, genomic, virulence, serotype, mobilome and epidemiological data. Different levels of access may be granted, but data modification is strictly reserved to the curators. It includes isolate identification, host, provider, date of isolation, geographical origin, phenotypic data, DNA extraction details, NGS information and genome assembly statistics. SalFoS currently contains NGS data and unpublished draft genomes from produce, human, animal and environmental isolates. Upon publication, draft genomes of SalFoS will become available at NCBI and Enterobase⁴.

⁴<http://enterobase.warwick.ac.uk/>

The SalFoS collection currently contains 2,498 entries for *Salmonella*, as well as for *Citrobacter*, *Hafnia* and *Proteus*, three genera often identified as false-positives by a number of *Salmonella* detection schemes. It includes previously described collections such as the unique *Salmonella* Genetic Stock Centre strains, described at <http://people.ucalgary.ca/~{}kesander/>. This collection was assembled with the aim of representing maximal genomic diversity.

SEQUENCING 4,500 *Salmonella* GENOMES: OBJECTIVES AND STRATEGY

Our working hypothesis is that a very high-quality, large-scale bacterial genome database available through a user-friendly pipeline will have a major impact for epidemiology, diagnosis, prevention and treatment. By generating a comprehensive genome sequence database truly representative of the foodborne *Salmonella* population, we will: (1) assemble a large and representative strain collection, with associated genome data, useful for antimicrobial testing, identification of resistance markers, data mining for new therapeutic targets and development of machine learning strategies; (2) develop



platforms and pipelines to manage and analyze this information, which will allow identification of prognostic markers, fast epidemiological tracking and reduction of socio-economic costs. We seek to develop user-friendly tools that will enable epidemiologists, microbiologists, clinicians and others to interpret genomic data, thus leading to informed decisions in cases of food contamination and outbreaks. The contamination of fresh produce by *Salmonella* will be addressed through the development of natural solutions to control the presence of *Salmonella* on fruits and vegetables as they are growing in the fields. New tests will also be developed so that fresh produce can be quickly and efficiently tested for the presence of *Salmonella* before being sold to consumers. In the context of outbreak investigation, the genomic data will be used to assess high-quality SNPs and core/whole genome MLST for their usefulness in genetic discrimination in addition to other emerging methods such as CRISPR and prophage sequence typing. As for outbreak investigation software, the National Microbiology Laboratory-Public Health Agency of Canada group has implemented the Integrated Rapid Infectious Disease Analysis project (IRIDA)⁵ and developed the SNVPhyl phylogenomics pipeline that is in use by PulseNet Canada for microbial genomic epidemiology (Petkau et al., 2016). A complementary system called the Metagenomics Computation and Analytics Workbench (MCAW) is being implemented as a computing service for food safety (Edlund et al., 2016; Weimer et al., 2016).

Sequencing for this project is performed on an Illumina MiSeq instrument (at the Plateforme d'Analyses Génomiques of the IBIS, Université Laval, Quebec City, QC, Canada), at a rate of 120 genomes per week, using 300 bp paired-end libraries, and with a median coverage of 45 \times . In order to perform core genome phylogenetic analysis, the pan-genome, i.e., the complete repertoire of genes of a species, is determined using a recently developed software capable of handling high-quality NGS data from thousands of genomes: Saturn V version 1.0⁶ (Jeukens et al., 2017). Additional analyses focus on genes implicated in virulence using comparative genomics predictions of confirmed and predicted virulence factors (Yang et al., 2008), and resistome identification based on the comprehensive antibiotic resistance database (CARD) (McArthur et al., 2013; Jia et al., 2016). A set of new reference *Salmonella* genomes representing maximal genomic diversity among foodborne pathogens will then be selected for PacBio Sequel sequencing to become fully assembled and annotated as a single circular chromosome.

THE IBIS BIOINFORMATICS PIPELINE FOR GENOME ASSEMBLY

When working with hundreds or thousands of genomes, analysis software for assembly, annotation, statistics for quality control and selection of additional reference genomes is required to extract relevant information in an automated and reliable fashion

with minimal human intervention. Ideally, this software should be platform independent and able to analyze sequence data directly without being tied to proprietary data formats. This insures maximal flexibility and reduces lag time to a minimum. We are currently using an integrated pipeline for *de novo* assembly of microbial genomes based on the A5 pipeline (Tritt et al., 2012). It was parallelized on a Silicon Graphics UV 300 using up to 120 cores to accommodate raw data from 120 genomes and provide assembly statistics as well as reference genome alignment metrics in as little as 2 h. This automated approach currently results in a median of 35 scaffolds per genome (median N50 = 462 kb).

PHYLOGENY OF *Salmonella*

Once isolates from a given outbreak are sequenced, patterns of shared variations can be used to infer which isolates within the outbreak are most closely related to each other (e.g., Didelot et al., 2017). As a future strategy for the Syst-OMICS project, this could be applied to partially sampled and on-going *Salmonella* outbreaks. Here, as a first step in the study of *S. enterica* diversity and epidemiology, we used 3,377 genomes; 1,627 were from a collaboration with UC Davis (Bart C. Weimer), and 1,750 were part of SalFoS. All genomes with >100 scaffolds were eliminated; this filter typically removes the vast majority of low coverage (i.e., low quality) assemblies and mixed cultures. As our assembly pipeline also includes alignment on a suite of reference genomes, it is also possible to ensure that genomes used belong to *S. enterica*. The core (conserved) genome was identified with Saturn V, and consisted of 839 genes, which were used for phylogenetic analysis. This number of core genes, which seems small compared to other studies (2,882 core genes for 73 genomes from 2 subspecies, Leekitcharoenphon et al., 2012), is due to both the extensive diversity and the high number of genomes used. As depicted in Figure 1, this population of *S. enterica* strains could be divided into seven major groups. They correspond to *S. enterica* subspecies *enterica* clades A and B and a collection of branching subspecies previously defined as *salamae*, *arizonae*, *diarizoneae*, *houtenae* and *indica*. The significant number of strains (3,377) included in our analysis and their wide-ranging sources (including environmental, human, animal and food) is essential to understand the diversity of *Salmonella* as a foodborne pathogen and in defining levels of virulence. The remarkable genomic diversity exhibited in Figure 1 is thought to enable the colonization of a wide range of ecological niches. The *Salmonella* Syst-OMICS consortium will provide fine-scale analysis of this diversity via virulence factors, antibiotic resistance genes as well as complete core and accessory genomes.

LINKING SalFoS WITH THE COMPREHENSIVE ANTIBIOTIC RESISTANCE DATABASE

The SalFoS database is intended to become an established platform for searching and comparing multiple genome

⁵<http://dev.irida.ca/>

⁶<https://github.com/ejfresch/saturnV>

sequences for *Salmonella* isolates. The database will also incorporate genome annotation and serotype prediction based on SISTR (Yoshida et al., 2016). Close attention to the links between specific genomic islands and patterns of SNPs in the core genome will help identify diagnostic sequences and SNP combinations for the development of new *Salmonella* subtyping methods with the highest resolution to date. This will be done using *de novo* island prediction with IslandViewer (Langille and Brinkman, 2009; Dhillon et al., 2015) as well as with gene presence-absence from SaturnV.

As an additional feature, we routinely determine the resistome of the genomes in SalFoS, i.e., the genes and variants likely involved in antibiotic resistance. This is done using the Resistance Gene Identifier (RGI) available with the CARD (McArthur et al., 2013; Jia et al., 2016), at <http://arpcard.mcmaster.ca/>. **Figure 2** summarizes the resistomes of 3,377 genomes. In fact, the original dataset contained 1,003 unique resistomes, composed of various combinations of 195 different genes and variants. Despite this impressive diversity, the most striking feature shown in **Figure 2** is that the two most frequently observed resistomes, which are extremely similar, account for 23% of the strains. They are therefore highly conserved and warrant further investigation. These results will be exploited to study and understand the pool of resistance genes present in *Salmonella* strains, with a focus on strains found in fresh produce, to understand the links between foodborne *Salmonella* and environmental strains with respect to resistance genes.

LINKING GENOMIC AND CLINICAL DATA

It will be essential to match phenotypic, epidemiological and available clinical *Salmonella* data (antibiotic resistance, virulence, and anonymized clinical observations) to the genomic data produced. We will categorize metadata in SalFoS so that isolates can be sorted by phenotype, allowing rapid identification of linked genomic signatures and the development of prognostic approaches for diagnostic, epidemiology and surveillance. We will develop tools to rapidly collate data for a given strain type and produce a concise phenotypic and clinical profile that provides users with an evidence-based decision-making platform. The Canadian Food Inspection Agency, Health Canada, Agriculture Canada, provincial public health laboratories and the National Microbiology Laboratory-Public Health Agency of Canada group are expected to be end-users of the projects outcomes.

REFERENCES

- Ashton, P. M., Nair, S., Peters, T. M., Bale, J. A., Powell, D. G., Painset, A., et al. (2016). Identification of *Salmonella* for public health surveillance using whole genome sequencing. *PeerJ* 4:e1752. doi: 10.7717/peerj.1752
- Bekal, S., Berry, C., Reimer, A. R., Van Domselaar, G., Beaudry, G., Fournier, E., et al. (2016). Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent *Salmonella enterica* serovar heidelberg clone in the context of outbreak investigations. *J. Clin. Microbiol.* 54, 289–295. doi: 10.1128/JCM.02200-15

FUTURE GENOMIC AND BIOLOGICAL STUDIES OF *Salmonella*

We will continuously improve SalFoS by adding *Salmonella* strains, NGS data and analysis as well as experimental results. Another aim of the Syst-OMICS consortium is to avoid duplication of efforts in *Salmonella* genomics and enhance interest from researchers having common goals. Additional members are welcome to join in and expand on our original Genome Canada project. We also intend to seek collaboration with other groups to connect our database with those developed for other *Salmonella* genomes. Finally, the *Salmonella* Syst-OMICS project could be a model for other groups interested in the bacterial genomics of infectious diseases, a strategy that we are also pursuing for *Pseudomonas aeruginosa* (Freschi et al., 2015).

AUTHOR CONTRIBUTIONS

J-GER, JJ, LF, IK-I and RL collected strains, performed the analyses and drafted the manuscript. BB provided support for sequencing and analysis. MD contributed to the development of SalFoS. All other authors handled strains and collected metadata. All authors revised the manuscript.

ACKNOWLEDGMENTS

We express our gratitude to members of the genomics analysis and bioinformatics platforms at IBIS. We also acknowledge Betty Wilkie, Ketna Mistry, Robert Holtslander and Shaun Kenaghan from the NML *Salmonella* reference laboratory for their assistance with serotyping. RL, LG, AG, ST, PD, DM, SG, SB, FD, SW, SM, GL, INF, YJ, PT, CN, RG, JoR, JW are funded by Genome Canada, provincial genome centers Génome Québec and Genome BC, and the Ontario Ministry of Research and Innovation. SM holds a Tier 1 Canada Research Chair in Bacteriophages.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.00996/full#supplementary-material>

- Casadevall, A. (2017). The pathogenic potential of a microbe. *mSphere* 2:e00015-17. doi: 10.1128/mSphere.00015-17
- Chen, L., Xiong, Z., Sun, L., Yang, J., and Jin, Q. (2012). VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* 40, D641–D645. doi: 10.1093/nar/gkr989
- D'costa, V. M., McGrann, K. M., Hughes, D. W., and Wright, G. D. (2006). Sampling the antibiotic resistome. *Science* 311, 374–377. doi: 10.1126/science.1120800
- Dhillon, B. K., Laird, M. R., Shay, J. A., Winsor, G. L., Lo, R., Nizam, F., et al. (2015). IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Res.* 43, 27. doi: 10.1093/nar/gkv401

- Didelot, X., Fraser, C., Gardy, J., and Colijn, C. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* 34, 997–1007. doi: 10.1093/molbev/msw275
- Edlund, S. B., Beck, K. L., Haiminen, N., Parida, L. P., Storey, D. B., Weimer, B. C., et al. (2016). Design of the MCAW compute service for food safety bioinformatics. *IBM J. Res. Dev.* 60:12. doi: 10.1147/JRD.2016.2584798
- Fatica, M. K., and Schneider, K. R. (2011). *Salmonella* and produce: survival in the plant environment and implications in food safety. *Virulence* 2, 573–579. doi: 10.4161/viru.2.6.17880
- Forest, C., Faucher, S. P., Poirier, K., Houle, S., Dozois, C. M., and Daigle, F. (2007). Contribution of the stg fimbrial operon of *Salmonella enterica* serovar typhi during interaction with human cells. *Infect. Immun.* 75, 5264–5271. doi: 10.1128/iai.00674-07
- Freschi, L., Jeukens, J., Kukavica-Ibrulj, I., Boyle, B., Dupont, M. J., Laroche, J., et al. (2015). Clinical utilization of genomics data produced by the international *Pseudomonas aeruginosa* consortium. *Front. Microbiol.* 6:1036. doi: 10.3389/fmicb.2015.01036
- Gal-Mor, O., Boyle, E. C., and Grassl, G. A. (2014). Same species, different diseases: how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. *Front. Microbiol.* 5:391. doi: 10.3389/fmicb.2014.00391
- Hoffmann, M., Luo, Y., Monday, S. R., Gonzalez-Escalona, N., Ottesen, A. R., Muruvanda, T., et al. (2016). Tracing origins of the *Salmonella* bareilly strain causing a food-borne outbreak in the United States. *J. Infect. Dis.* 213, 502–508. doi: 10.1093/infdis/jiv297
- Hoffmann, M., Zhao, S., Pettengill, J., Luo, Y., Monday, S. R., Abbott, J., et al. (2014). Comparative genomic analysis and virulence differences in closely related *Salmonella enterica* serotype Heidelberg isolates from humans, retail meats, and animals. *Genome Biol. Evol.* 6, 1046–1068. doi: 10.1093/gbe/evu079
- Jackson, B. R., Griffin, P. M., Cole, D., Walsh, K. A., and Chai, S. J. (2013). Outbreak-associated *Salmonella enterica* serotypes and food commodities, United States, 1998–2008. *Emerg. Infect. Dis.* 19, 1239–1244. doi: 10.3201/eid1908.121511
- Jeukens, J., Freschi, L., Vincent, A. T., Emond-Rheault, J. G., Kukavica-Ibrulj, I., Charette, S. J., et al. (2017). A pan-genomic approach to understand the basis of host adaptation in *Achromobacter*. *Genome Biol. Evol.* doi: 10.1093/gbe/evx061 [Epub ahead of print].
- Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., et al. (2016). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 45, D566–D573. doi: 10.1093/nar/gkw1004
- Kheadr, E., Zihler, A., Dabour, N., Lacroix, C., Le Blay, G., and Fliss, I. (2010). Study of the physicochemical and biological stability of pediocin PA-1 in the upper gastrointestinal tract conditions using a dynamic in vitro model. *J. Appl. Microbiol.* 109, 54–64. doi: 10.1111/j.1365-2672.2009.04644.x
- Kim, S. (2010). *Salmonella* serovars from foodborne and waterborne diseases in Korea, 1998–2007: total isolates decreasing versus rare serovars emerging. *J. Kor. Med. Sci.* 25, 1693–1699. doi: 10.3346/jkms.2010.25.12.1693
- Kozak, G. K., Macdonald, D., Landry, L., and Farber, J. M. (2013). Foodborne outbreaks in Canada linked to produce: 2001 through 2009. *J. Food Prot.* 76, 173–183. doi: 10.4315/0362-028X.JFP-12-126
- Langille, M. G. I., and Brinkman, F. S. L. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 25, 664–665. doi: 10.1093/bioinformatics/btp030
- Lanza, V. F., De Toro, M., Garcillan-Barcia, M. P., Mora, A., Blanco, J., Coque, T. M., et al. (2014). Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genet.* 10:e1004766. doi: 10.1371/journal.pgen.1004766
- Le Blay, G., Hammami, R., Lacroix, C., and Fliss, I. (2012). Stability and inhibitory activity of pediocin PA-1 against *Listeria* sp. in simulated physiological conditions of the human terminal ileum. *Probiotics Antimicrob. Proteins* 4, 250–258. doi: 10.1007/s12602-012-9111-1
- Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aarestrup, F. M., and Ussery, D. W. (2012). Genomic variation in *Salmonella enterica* core genes for epidemiological typing. *BMC Genomics* 13:88. doi: 10.1186/1471-2164-13-88
- Little, T. J., Allen, J. E., Babayan, S. A., Matthews, K. R., and Colegrave, N. (2012). Harnessing evolutionary biology to combat infectious disease. *Nat. Med.* 18, 217–220. doi: 10.1038/nm.2572
- Liu, Y. Y., Chen, C. C., and Chiou, C. S. (2016). Construction of a pan-genome allele database of *Salmonella enterica* serovar enteritidis for molecular subtyping and disease cluster identification. *Front. Microbiol.* 7:2010. doi: 10.3389/fmicb.2016.02010
- Majowicz, S. E., Musto, J., Scallan, E., Angulo, F. J., Kirk, M., O'Brien, S. J., et al. (2010). The global burden of nontyphoidal *Salmonella* gastroenteritis. *Clin. Infect. Dis.* 50, 882–889. doi: 10.1086/650733
- Mcarthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., et al. (2013). The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* 57, 3348–3357. doi: 10.1128/AAC.00419-13
- Nuesch-Inderbinen, M., Cernela, N., Althaus, D., Hachler, H., and Stephan, R. (2015). *Salmonella enterica* serovar szentes, a rare serotype causing a 9-month outbreak in 2013 and 2014 in switzerland. *Foodborne Pathog. Dis.* 12, 887–890. doi: 10.1089/fpd.2015.1996
- Perry, J. A., and Wright, G. D. (2014). Forces shaping the antibiotic resistome. *Bioessays* 36, 1179–1184. doi: 10.1002/bies.201400128
- Petkau, A., Mabon, P., Sieffert, C., Knox, N., Cabral, J., Iskander, M., et al. (2016). SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *bioRxiv* 092940. doi: 10.1101/092940
- Podolak, R., Enache, E., Stone, W., Black, D. G., and Elliott, P. H. (2010). Sources and risk factors for contamination, survival, persistence, and heat resistance of *Salmonella* in low-moisture foods. *J. Food Prot.* 73, 1919–1936.
- Ribot, E. M., and Hise, K. B. (2016). Future challenges for tracking foodborne diseases: pulseNet, a 20-year-old US surveillance system for foodborne diseases, is expanding both globally and technologically. *EMBO Rep.* 17, 1499–1505. doi: 10.1525/embr.201643128
- Rouli, L., Merhej, V., Fournier, P. E., and Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* 7, 72–85. doi: 10.1016/j.nmni.2015.06.005
- Roy, M.-F., Riendeau, N., Bédard, C., Hélie, P., Min-Oo, G., Turcotte, K., et al. (2007). Pyruvate kinase deficiency confers susceptibility to *Salmonella* Typhimurium infection in mice. *J. Exp. Med.* 204, 2949–2961. doi: 10.1084/jem.20062606
- Scharff, R. L., Besser, J., Sharp, D. J., Jones, T. F., Peter, G.-S., and Hedberg, C. W. (2016). An Economic Evaluation of PulseNet. *Am. J. Prev. Med.* 50, S66–S73. doi: 10.1016/j.amepre.2015.09.018
- Shariat, N., and Dudley, E. G. (2014). CRISPRs: molecular signatures used for pathogen subtyping. *Appl. Environ. Microbiol.* 80, 430–439. doi: 10.1128/AEM.02790-13
- Thomas, M. K., Murray, R., Flockhart, L., Pintar, K., Fazil, A., Nesbitt, A., et al. (2015). Estimates of foodborne illness-related hospitalizations and deaths in Canada for 30 specified pathogens and unspecified agents. *Foodborne Pathog. Dis.* 12, 820–827. doi: 10.1089/fpd.2015.1966
- Timme, R. E., Pettengill, J. B., Allard, M. W., Strain, E., Barrangou, R., Wehnert, C., et al. (2013). Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. enterica inferred from genome-wide reference-free SNP characters. *Genome Biol. Evol.* 5, 2109–2123. doi: 10.1093/gbe/evt159
- Tritt, A., Eisen, J. A., Facciotti, M. T., and Darling, A. E. (2012). An integrated pipeline for de novo assembly of microbial genomes. *PLoS ONE* 7:e42304. doi: 10.1371/journal.pone.0042304
- Weedmark, K. A., Mabon, P., Hayden, K. L., Lambert, D., Van Domselaar, G., Austin, J. W., et al. (2015). Clostridium botulinum Group II isolate phylogenomic profiling using whole-genome sequence data. *Appl. Environ. Microbiol.* 81, 5938–5948. doi: 10.1128/aem.01155-15
- Weimer, B. C., Storey, D. B., Elkins, C. A., Baker, R. C., Markwell, P., Chambliss, D. D., et al. (2016). Defining the food microbiome for authentication, safety, and process management. *IBM J. Res. Dev.* 60:13. doi: 10.1147/JRD.2016.2582598
- Yang, J., Chen, L., Sun, L., Yu, J., and Jin, Q. (2008). VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res.* 36, D539–D542. doi: 10.1093/nar/gkm951
- Yoshida, C. E., Kruczakiewicz, P., Laing, C. R., Lingohr, E. J., Gannon, V. P., Nash, J. H., et al. (2016). The *Salmonella* in silico typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS ONE* 11:e0147101. doi: 10.1371/journal.pone.0147101

Conflict of Interest Statement: The handling Editor declared a shared affiliation, though no other collaboration, with the authors ST and AG, and the handling Editor states that the process met the standards of a fair and objective review.

The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Emond-Rheault, Jeukens, Freschi, Kukavica-Ibrulj, Boyle, Dupont, Colavecchio, Barrere, Cadieux, Arya, Bekal, Berry, Burnett, Cavestri, Chapin, Crouse, Daigle, Danyluk, Delaquis, Dewar, Doualla-Bell, Fliss, Fong, Fournier,

Franz, Garduno, Gill, Gruenheid, Harris, Huang, Huang, Johnson, Joly, Kerhoas, Kong, Lapointe, Larivière, Loignon, Malo, Moineau, Mottawea, Mukhopadhyay, Nadon, Nash, Ngueng Feze, Ogunremi, Perets, Pilar, Reimer, Robertson, Rohde, Sanderson, Song, Stephan, Tamber, Thomassin, Tremblay, Usongo, Vincent, Wang, Weadge, Wiedmann, Wijnands, Wilson, Wittum, Yoshida, Youfsi, Zhu, Weimer, Goodridge and Levesque. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The Validation and Implications of Using Whole Genome Sequencing as a Replacement for Traditional Serotyping for a National *Salmonella* Reference Laboratory

Chris A. Yachison^{1,2}, Catherine Yoshida³, James Robertson³, John H. E. Nash³, Peter Kruczakiewicz⁴, Eduardo N. Taboada⁴, Matthew Walker¹, Aleisha Reimer¹, Sara Christianson¹, Anil Nichani³, The PulseNet Canada Steering Committee and Celine Nadon^{1,2*}

OPEN ACCESS

Edited by:

Sandra Torriani,
University of Verona, Italy

Reviewed by:

David Rodriguez-Lazaro,
University of Burgos, Spain
Alejandro Castillo,
Texas A&M University, United States

*Correspondence:

Celine Nadon
celine.nadon@phac-aspc.gc.ca

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 30 November 2016

Accepted: 24 May 2017

Published: 09 June 2017

Citation:

Yachison CA, Yoshida C, Robertson J, Nash JHE, Kruczakiewicz P, Taboada EN, Walker M, Reimer A, Christianson S, Nichani A, The PulseNet Canada Steering Committee and Nadon C (2017) The Validation and Implications of Using Whole Genome Sequencing as a Replacement for Traditional Serotyping for a National *Salmonella* Reference Laboratory. *Front. Microbiol.* 8:1044. doi: 10.3389/fmicb.2017.01044

Salmonella serotyping remains the gold-standard tool for the classification of *Salmonella* isolates and forms the basis of Canada's national surveillance program for this priority foodborne pathogen. Public health officials have been increasingly looking toward whole genome sequencing (WGS) to provide a large set of data from which all the relevant information about an isolate can be mined. However, rigorous validation and careful consideration of potential implications in the replacement of traditional surveillance methodologies with WGS data analysis tools is needed. Two *in silico* tools for *Salmonella* serotyping have been developed, the *Salmonella* *in silico* Typing Resource (SISTR) and SeqSero, while seven gene MLST for serovar prediction can be adapted for *in silico* analysis. All three analysis methods were assessed and compared to traditional serotyping techniques using a set of 813 verified clinical and laboratory isolates, including 492 Canadian clinical isolates and 321 isolates of human and non-human sources. Successful results were obtained for 94.8, 88.2, and 88.3% of the isolates tested using SISTR, SeqSero, and MLST, respectively, indicating all would be suitable for maintaining historical records, surveillance systems, and communication structures currently in place and the choice of the platform used will ultimately depend on the users need. Results also pointed to the need to reframe serotyping in the genomic era as a test to understand the genes that are carried by an isolate, one which is not necessarily congruent with what is antigenically expressed. The adoption of WGS for serotyping will provide the simultaneous collection of information that can be used by multiple programs within the current surveillance paradigm; however, this does not negate the importance of the various programs or the role of serotyping going forward.

Keywords: *Salmonella enterica*, whole genome sequencing, serotyping, serovars, surveillance, phenotype prediction

INTRODUCTION

Salmonella enterica is one of the most prevalent foodborne pathogens world-wide and a major causative agent of gastroenteritis in North America (Kim et al., 2006; Hendriksen et al., 2011; Parmley et al., 2013; Thomas et al., 2013; World Health Organization, 2015). Within Canada, *S. enterica* is estimated to be responsible for over 87,000 cases of domestically acquired foodborne illnesses each year (Thomas et al., 2013), leading to 11,600 hospitalizations and 238 deaths (Thomas et al., 2015). In Canada, the surveillance of this priority pathogen is a multifaceted endeavor, encompassing multiple surveillance programs that monitor *Salmonella* along the farm to fork to human clinical disease spectrum (Parmley et al., 2013). While the individual mandates of the various surveillance systems differ (Swaminathan et al., 2006; Government of Canada, 2014), all heavily rely on the classification of isolates into serovars by the globally recognized serotyping system (Parmley et al., 2013); the White-Kauffmann-Le Minor (WKL) scheme (Grimont and Weill, 2007).

The *Salmonella* serotyping system classifies isolates on the basis of the immunological reactions to cell surface antigens, specifically the somatic O and the two variably expressed flagellar H antigens, denoted H1 and H2 (Kim et al., 2006; Wattiau et al., 2011; Zhang et al., 2015; Yoshida et al., 2016b). The most recent edition of the WKL scheme has identified over 2,500 serovars belonging to the five subspecies of *S. enterica* (Grimont and Weill, 2007). However, it is important to note that the majority of human clinical disease is the result of a select few important human pathogenic serovars (Hendriksen et al., 2011). Traditionally, serotyping is performed through the phenotypic characterization of the O and H antigens via the slide agglutination test, in which the clumping of cells is observed in response to specific antisera. Although this technique is widely used (Wattiau et al., 2011), it can be time consuming and laborious (Kim et al., 2006; Wattiau et al., 2011; Zhang et al., 2015), and still leave five to eight percent of all *Salmonella* isolates untypeable (Kim et al., 2006).

Increasingly, the global food-safety and public health communities are looking toward the ‘omics’ technologies to provide a rapid, cost effective, high throughput, and expansive set of data, from which all the relevant information about an isolate or organism can be mined (Bergholz et al., 2014; Ronholm et al., 2016). The technological advancements of the whole genome sequencing (WGS) platforms, such as the Illumina MiSeq, and the improved bioinformatic analyses are revolutionizing surveillance programs. WGS promises to not only provide the best dataset available to determine pathogen relatedness, fulfilling a key mandate of programs such as PulseNet Canada, but WGS data could also be used to derive information about other important characteristics, such as serotype, using *in silico* workflows (Gilmour et al., 2013; Ronholm et al., 2016).

In recent years, multiple applications for the *in silico* determination of *Salmonella* serovars from WGS data have been developed (Achtman et al., 2012; Zhang et al., 2015; Yoshida et al., 2016b). Achtman et al. (2012) suggested that the serovar of an isolate could be inferred from multi-locus sequence typing

(MLST) (Achtman et al., 2012). MLST assigns unique alleles to different sequences across 400–500bp fragments of specified housekeeping genes (loci). The alleles assigned across all the loci are used to define the individual sequence types (ST), converting a large amount of data into an easily defined single number (Maiden et al., 1998). In the MLST serotyping scheme developed for *Salmonella* seven housekeeping gene fragments are typed and the individual ST or their larger clonal complexes are then linked to *Salmonella* serovars. While MLST is not strictly an *in silico* analysis, the MLST data can be mined directly from WGS data using bioinformatic pipelines (Achtman et al., 2012; Ashton et al., 2016). Applications such as SeqSero¹ and the *Salmonella* *in silico* Typing Resource (SISTR²) have utilized an approach in which the gene sequences encoding the individual somatic and flagellar antigens are used to determine an isolate’s serovar. SeqSero extracts the relevant genomic regions, specifically the *rfb* region for somatic antigen determination and the *fliC* and *fliB* genes for H1 and H2 antigen determination, from the genome assemblies and aligns these regions to curated databases (Zhang et al., 2015) using BLAST (Basic Local Alignment Search Tool) (Camacho et al., 2009). The overall antigenic formula is then looked up in the WKL, much like traditional serotyping (Zhang et al., 2015). The SISTR platform looks at the highly variable *wzx* and *wzy* genes of the *rfb* region to determine the somatic antigen, and the *fliC* and *fliB* genes to determine the H1 and H2 antigens, respectively. The predicted antigenic formula is queried against the WKL serovar database to provide a serovar name. Unlike flagellar antigens, the somatic O antigens are the result of complex molecular pathways and at present there is not enough known about the biology of the organism to make predictions of the individual O antigens expressed, but using the information present in the *wzx/wzy* genes it is possible to assign isolates to a serogroup. Additional testing may be required to resolve the ambiguous serovar designations as a result of the serogroup vs individual antigen determination. To resolve these ambiguities SISTR uses a novel 330 locus core genome MLST (cgMLST) analysis to provide a phylogenetic context; where the predominant serovar within the cgMLST cluster is used to choose the most likely serovar from the list of potential serovars. Together these two methodologies are used to provide an overall SISTR serovar prediction with improved concordance with traditional methods (Yoshida et al., 2016b).

Before WGS data can be used for the serotyping of *Salmonella*, there needs to be rigorous validation of any proposed method and careful implementation plans must be prepared to ensure the data is appropriate to replace the traditional serotyping method so there is minimal disruption to current surveillance programs. Validation of the *in silico* serotyping tools needs to be performed using a heavily curated panel of isolates with reliable identifications to ensure that the produced results are accurate (Gilmour et al., 2013). The objective of this project was to: validate and compare *in silico* serotyping methodologies, MLST, SeqSero, and SISTR; assess the impact of *in silico* serotyping on existing surveillance systems; and develop the considerations that

¹<http://www.denglab.info/SeqSero>

²<https://lfz.corefacility.ca/sistr-app/>

can guide the future implementation of *in silico* serotyping for national surveillance in Canada and elsewhere.

MATERIAL AND METHODS

Study Design

A total of 813 *Salmonella enterica* isolates were selected for this study to assess the performance of the three *in silico* serotyping methods, SISTR, SeqSero, and MLST in comparison to phenotypic serotyping. Results from the validation were used to identify the implications of using an *in silico* serotyping tool and develop considerations for the replacement of traditional serotyping with said tool.

Study Isolates and Traditional Serotyping

All isolates used in this study were previously serotyped using traditional methods as follows. Isolates were grown overnight at 37°C on Luria-Bertani agar (BD Canada, Mississauga, ON, Canada) and the antigenic formula of each strain was determined using standard methods (Shipp and Rowe, 1980; Ewing, 1986), with the serovar assigned according to the WKL scheme in an OIE/ISO accredited laboratory (Grimont and Weill, 2007). Isolates were all phenotypically serotyped in the Reference Laboratory at the National Microbiology Laboratory (NML).

The 813 isolates used in this study could be split into two groups on the basis of their rationale for inclusion in the study. Group one consisted of 492 isolates from multiple target serovars that were chosen for their clinical relevance and importance in diagnostic and reference laboratories serving surveillance functions in Canada. Four-hundred of the 492 isolates from this group were split evenly among the top twenty serovars reported to NESP in 2012. Together these serovars represent about 85% of all reported cases of human salmonellosis in Canada (Government of Canada, 2014), and include serovars: Enteritidis, Heidelberg, Typhimurium, ssp I 4,[5],12:i:-, Thompson, Infantis, Newport, Typhi, ssp I 4,[5],12:b:-, Braenderup, Saintpaul, Javiana, Paratyphi A, Hadar, Agona, Paratyphi B var. Java, Stanley, Oranienburg, Muenchen, and Montevideo. The remaining 92 isolates from this group represented a collection of clinically relevant but infrequently encountered serovars in Canada. These 92 isolates included: serovars of increased clinical importance, either due to increased association with invasive or travel related infections (specifically serovars Dublin, Cerro, Schwarzengrund, Sandiego, Panama, and Corvallis); serovars deemed difficult to differentiate by traditional serotyping (specifically serovars Senftenberg and Kouka; Carrau and Madelia; Lattenkamp; and Paratyphi B); serovars from non-subspecies I (specifically from subspecies II, IIIa, IIIb, and IV); and isolates left untypeable by traditional serotyping. All isolates collected as part of this group were randomly selected from the total population of *Salmonella* isolates from their respective serovars that were submitted to the NML in Winnipeg between the years 2009–2013. Group two consisted of 321 isolates chosen to represent the most globally prevalent *Salmonella* serovars from both human

and non-human sources. The isolates from this group were previously used in a validation study evaluating other molecular typing methods and were collected from human, animal, and environmental sources (Yoshida et al., 2016a). One hundred and fifty-two isolates in this grouping were from the target serovars outlined above, while the other 169 isolates in this group were from a collection of other serovars, meant to provide coverage for the majority of antigenic determinants currently described.

Genome Sequencing and Assembly

Genomic DNA was extracted from group one isolates using overnight cultures grown in LB-Lennox 0.5% NaCl broth via the Qiagen DNeasy 96 Blood and Tissue Kit (Qiagen Ltd., Mississauga, ON, Canada) or from a nutrient agar plate using the Epicenter Metagenomics DNA isolation kit for water (Mandel Scientific Company Inc., Guelph, ON, Canada). For isolates from group two, genomic DNA was extracted from isolated colonies on overnight Luria_Bertani agar plates using the KingFisher Cell and Tissue DNA Kit (VWR, Mississauga, ON, Canada) on the KingFisher Flex (VWR) or using the EZ1 DNA tissue kit and BioRobot (Qiagen). Manufacturer's instructions were followed with the addition of 100 g of lysozyme (Sigma-Aldrich Canada Ltd., Oakville, ON, Canada; 10 mg/ml) in the cell lysis incubation stage for isolates from group two.

Recovered DNA for all isolates in the study was quantified with a Qubit DNA quantification system (Invitrogen Canada Inc., Burlington, ON, Canada) and diluted down to a genomic DNA concentration of 0.2 ng/μl. Sample libraries for all isolates were prepared using the MiSeq Nextera XT library preparation kit (Illumina, Inc., San Diego, CA, United States). Libraries were size selected for a minimum insert size of 500bp using the BluePippin (Sage Science, Beverly, MA, United States). Paired end sequencing was performed on the Illumina MiSeq with the MiSeq Reagent Kit v3 600 cycles (2 × 300bp forward and reverse) to achieve an average estimated genome coverage greater than 30× for all isolates. Raw sequencing read data was uploaded to NCBI for all isolates. Group one isolates have been uploaded under BioProject PRJNA353625, while isolates from group two have been uploaded under BioProject PRJNA354244.

The paired end reads were first merged using FLASH (version 1.2.9³) (Magoc and Salzberg, 2011) and then *de novo* assembled into contigs using SPAdes (version 3.5.0⁴) (Bankevich et al., 2012). SPAdes was run using the careful option to correct assembly errors and the resulting FASTA files were used in downstream analysis.

Interpretation and Scoring of Results

All isolates were uploaded to the SISTR website⁵ (Yoshida et al., 2016b) and SeqSero website⁶ (Zhang et al., 2015) for serovar prediction. An *in silico* 7-gene MLST ST was generated using the SISTR platform and ST data from the platform was compared to

³<http://ccb.jhu.edu/software/FLASH>

⁴<http://bioinf.spbau.ru/spades>

⁵<https://lfz.corefacility.ca/sistr-app/>

⁶<http://www.denglab.info/SeqSero>

the University of Warwick *Salmonella enterica* MLST database⁷ to generate the MLST serovar prediction (Achtman et al., 2012). A comparison of results to traditional serotyping was done using the interpretation criteria as described below. The results were categorized into full, inconclusive, incongruent or incorrect matches. Matches were considered “full” when the overall serovar prediction was concordant with the reported serovar by traditional typing. Matches were considered “inconclusive” when the overall serovar prediction was ambiguous (multiple serovars indicated), or of partial prediction (information was missing in overall prediction, but the individual parts provided were correct). Matches were considered “incongruent” when the overall serovar prediction was incongruent with the reported phenotypic serovar due to the carriage of antigenic determinants (either phase two flagellar or somatic antigen genes) that were not expressed phenotypically. Matches were considered “incorrect” when the overall serovar prediction was incorrect with respect to the reported serovar by traditional serotyping. Successful predictions were calculated based on the proportion of results that were categorized as full, inconclusive, or incongruent matches, indicating a positive test result in relation to traditional serotyping.

A Fisher’s exact test was used to evaluate the statistical significance between the successful predictions from each platform using a 2-by-2 contingency table via Graphpad Quickcalcs⁸. A *P*-value of less than or equal to 0.05 was considered statistically significant.

The test sensitivity and specificity for the prediction of Enteritidis and Typhimurium was also assessed for each platform in comparison to standard serotyping through a 2-by-2 table analysis (Mackinnon, 2000), following the removal of any incongruent results from the analyses. Incongruent results were removed from the analysis due to their inherent incompatibility between the phenotypically and genetically derived test results which could artificially reduce the sensitivity and specificity measured. Test sensitivity was defined as $TP/(TP+FN)$ and test specificity was defined as $TN/(TN+FP)$, where TP = true positives, FN = false negative, TN = true negatives, and FP = false positives.

RESULTS

The performance of the three *in silico* methods for *Salmonella* serovar prediction was assessed using a panel of 813 isolates, including 492 Canadian clinical isolates and 321 isolates of human and non-human sources. The three methods, SISTR, SeqSero, and MLST provided successful results for 94.8, 88.2, and 88.3% of the 813 isolates tested, respectively; where successful results were considered when the result did not include incorrect information. These successful identifications were further broken down into full matches, inconclusive matches, and incongruent matches. Full matches were recorded when there was an identical

serovar match between traditional serotyping and the *in silico* method under study. We recorded full matches in 89.7, 54.1, and 77.9% of the isolates tested using SISTR, SeqSero, and MLST, respectively. Inconclusive matches were recorded when the information provided by the *in silico* typing method was insufficient and would require further laboratory analysis to determine or narrow down the correct serovar. Inconclusive matches were reported with SISTR, SeqSero, and MLST for 1.1, 30.0, and 6.4%, respectively, of the isolates tested. Incongruent matches were when the phenotypic serovar prediction was incongruent with the *in silico* prediction due to the carriage of unexpressed antigenic determinants. We noted incongruent matches in 4.1% of the isolates tested regardless of testing method. Lastly, incorrect results were reported in 5.2, 11.8, and 11.7% of the isolates tested using SISTR, SeqSero, and MLST, respectively. A summary of these results is presented in Table 1. A breakdown of the results for individual serovars within the target group is included in Supplementary Table 1. The number of successful results reported by SISTR was deemed to be significantly greater than the number of successful results reported by either SeqSero or MLST (one-tailed *p*-values of less than 0.001). However, the observed differences in number of successful results recorded by SeqSero and MLST was not deemed to be statistically significant.

All rough isolates were naturally considered to be incongruent matches, and a complete serovar call was generated for 96, 54, and 85% of the rough isolates tested by SISTR, SeqSero, and MLST analysis methods, respectively. While partial results were generated for all isolates tested using SISTR and SeqSero and for 85% of the isolates tested using MLST analysis. Predictions across the platforms were consistent with each other and any reported H antigens, except for one MLST predictions which were inconsistent with the reported H antigens, and one isolate whose H2 antigen prediction differed between SISTR and SeqSero, but could not be typed via traditional serotyping. No single genetic change or phylogenetic signal was detected amongst the 26 rough isolates tested as part of our study.

Test sensitivity and specificity was calculated for each *in silico* serotyping method for serovars Enteritidis and Typhimurium due to their increased global prevalence and importance. Test sensitivity for serovar Enteritidis was 95.2, 97.1, and 87.0% for the SISTR, SeqSero, and MLST methods, respectively; while test specificity for serovar Enteritidis was 99.7, 98.9, and 99.7%. SISTR, SeqSero, and MLST displayed test sensitivities of 92.9, 100, and 65% and test specificities of 100, 98.3, and 100% for serovar Typhimurium, respectively. A summary of the test specificity and test sensitivity results for these two serovars across the analysis methods is displayed in Table 2.

DISCUSSION

Two platforms have been developed for *in silico* *Salmonella* serovar prediction, SISTR (Yoshida et al., 2016b), and SeqSero (Zhang et al., 2015), while 7-gene MLST analysis (Achtman et al., 2012) can be adapted as an *in silico* serovar test (Ashton

⁷<http://mlst.warwick.ac.uk/mlst/>

⁸<http://www.graphpad.com/quickcalcs>

TABLE 1 | Performance of the three *in silico* methods for *Salmonella* serovar prediction, SISTR, SeqSero, and MLST, compared to traditional serotyping for 813 *Salmonella enterica* isolates.

Platform	Successful results			Total successful results	Incorrect results	Total tested
	Full Match	Inconclusive Match	Incongruent Match			
SISTR	729 (89.7%)	9 (1.1%)	33 (4.1%)	771* (94.8%)	42 (5.2%)	813
SeqSero	440 (54.1%)	244 (30.0%)	33 (4.1%)	717 (88.2%)	96 (11.8%)	813
MLST	633 (77.9%)	52 (6.4%)	33 (4.1%)	718 (88.3%)	95 (11.7%)	813

*Statistically significant improvement compared to other *in silico* platforms, one-tailed *p*-values ≤ 0.001 .

TABLE 2 | Sensitivity and specificity of prediction of *Salmonella* serovars Enteritidis and Typhimurium using three *in silico* methods for *Salmonella* serovar determination, SISTR, SeqSero, and MLST.

	Method	TP ¹	TN ²	FP ³	FN ⁴	Total tested	Sensitivity	Specificity
<i>Salmonella</i> serovar Enteritidis	SISTR	40	736	2	2	780	95.2	99.7
	SeqSero	34	737	1	8	780	97.1	98.9
	MLST	40	732	6	2	780	87.0	99.7
<i>Salmonella</i> serovar Typhimurium	SISTR	39	738	3	0	780	92.9	100
	SeqSero	26	741	0	13	780	100	98.3
	MLST	39	720	21	0	780	65.0	100

¹True positives; ²True negatives; ³False positives; ⁴False negatives.

et al., 2016) and all were assessed in the current study. We report that successful results were obtained for 94.8, 88.2, and 88.3% of the 813 *Salmonella* isolates tested using SISTR, SeqSero, and MLST, respectively (Table 1). These data are largely consistent with the previously reported success rate of each individual platform (Zhang et al., 2015; Ashton et al., 2016; Yoshida et al., 2016b). In our study SISTR significantly outperformed both SeqSero and MLST analysis for the *in silico* serotyping of isolates. Previous global surveys of quality assurance testing for traditional *Salmonella* serotyping methods have found participating laboratories worldwide were able to correctly identify an average of 82% of stains tested (Hendriksen et al., 2009), and this shows the overall suitability of replacing phenotypic serotyping with an *in silico* analysis tool. The test sensitivity and specificity of the platforms for serovars Enteritidis and Typhimurium was also assessed. The test sensitivity values ranged from a low of 65% for serovar Typhimurium using the MLST analysis platform to a high 100% for the same serovar using the SeqSero platform, with the rest of the values falling between 87 and 97.1%. While the test specificity values ranged from 98.3 to 100% across the platforms for the two serovars (Table 2). In all, these data indicate that all platforms show a high degree of accuracy for detecting the two most commonly reported serovars to public health laboratories worldwide (Hendriksen et al., 2011), except for some difficulty with false calls of serovar Typhimurium using the MLST platform, due to the grouping of Typhimurium and its monophasic variant (ssp I 4,[5],12:i:-) into a single category by this method (Achtman et al., 2012). Proper categorization of Enteritidis and Typhimurium is important for some jurisdictions due to strict and costly regulatory controls (Yoshida et al., 2014). While all platforms displayed a high overall success rate the breakdown of the successful results into more informative categories provides a better picture of the results from the platforms.

Implications of Findings

Inconclusive Results

All platforms reported some inconclusive results ranging from a low of 1.1% using SISTR to a high of 30% using SeqSero. Inconclusive results across the platforms were the result of ambiguous calls (multiple serovars listed), lack of serovar variant determination, or a complete lack of serotype reported. Isolates that returned with inconclusive results would require further analysis using the traditional serotyping techniques or other biochemical tests to provide a full serovar call. While these inconclusive results may be of little concern to some users, it may be concerning to others wanting a full answer. For laboratories transitioning away from traditional serotyping that still require a full result this could present a complication, and would require the maintenance of the technical and logistical capacity to carry out traditional serotyping or the submission to a reference laboratory where the generation of results may be delayed. SeqSero specifically displayed difficulty differentiating serovars that have the same antigenic formula but differed on minor O antigenic factors, such as Carrau (6,14,[24]:y:1,7) vs. Madelia (1,6,14,25:y:1,7), or that differed in subspeciation, such as Javiana (subspecies I 1,9,12:l,z28:1,5) vs. subspecies II 9,12:l,z28:1,5. However, there is some evidence that certain minor antigenic factors such as O:6 are variably expressed, which would mean that serovars such as Hadar (6,8:z10:e,n,x) and Istanbul (8:z10:e,n,x) are actually one and the same (Mikoleit et al., 2012). An updated WKL scheme could potentially resolve these issues thereby reducing the number of inconclusive matches recorded by SeqSero; as serovars with these antigenic formulas were reported as both possible options. Meanwhile, the ability of SISTR to resolve ambiguities in the antigenic calls of serovars and the ability of the MLST method to provide a serovar call is only as good as the databases from which they draw their phylogenetic connections to serovars. For rare and unusual serovars this can

lead to ambiguous results in SISTR or non-calls by the MLST method and in both these cases traditional serotyping would be required to provide a full serovar call.

Incongruent Results

The incongruent isolates uncovered in our validation point toward one of the major implications of using WGS analysis platforms for serotyping, specifically, that a genomic test is not a full equivalent to a phenotypic test in all cases. Using genomics to answer the serovar question requires us to adopt a new paradigm where the carriage of genes for O and H antigens determines the serovar, not the expression of these genes. While this new paradigm presents an opportunity for antigen determination of previously untypeable isolates it also leads to cases of incongruency between *in silico* serotyping tests and traditional methods, specifically in regards to some important monophasic serovars and isolates that are considered rough. For important monophasic serovars such as ssp I 4,[5],12:b:- and ssp I 4,[5],12:i:- previous research has shown that mutations in the flagellar phase variation machinery can lead to the loss of H2 antigen expression without the loss of the H2 antigenic gene determinant, *fljB* (Toboldt et al., 2013; Boland et al., 2015). Multiple mutations in this region have been reported including single base pair or full gene deletions (Toboldt et al., 2013) and insertions of IS26 elements (Boland et al., 2015). Of the 41 monophasic isolates tested in our panel, seven still carried the *fljB* gene and displayed some of the aforementioned mutations within their flagellar phase variation regions (data not shown). Meanwhile rough isolates have lost the ability to express O antigens (Reyes et al., 2012), and evidence suggests that the 'roughness' seen in traditional serotyping is not the result of a single genetic loss or change, but instead the result of various mutations, frameshifts or full gene deletions to various genes within the *rfb* region (Kong et al., 2011) which was confirmed through the analysis of our rough isolates.

As we move toward using WGS as a replacement for serotyping it is important that downstream consumers of this information accommodate for the fact that there may be minor changes in the prevalence of key serovars due to isolates possessing antigenic genes that are not expressed. It is important to note that the inconsistencies seen with regards to some monophasic and rough isolates are not the result of incorrect answers, but the result of different ways of framing the serotyping question. Understanding the intricacies of the phenotypic expression of antigenic markers is beyond our current scope of knowledge, and would require detailed studies on the effects of various mutations across a multitude of genes to determine their effect (Ronholm et al., 2016). As surveillance records are important documents for a multitude of users the identification of how the results were generated, version numbers of platforms used, and any limitations of the methods used is crucial so end users know how the results were generated.

Incorrect Results

All analysis platforms reported the presence of some incorrect results amongst the 813 isolates tested, ranging from a low of 5.2% of isolates tested using SISTR to just over 11% of isolates

tested using SeqSero and MLST. For SeqSero and SISTR some incorrect results were due to incorrect calling of various antigenic determinants, especially in regards to closely related serovars, such as those that differ on the basis of flagellar antigens of the g-complex. This is a limitation of both traditional serotyping (Hendriksen et al., 2009) and some molecular based serotyping techniques (Yoshida et al., 2014) and is related to the high sequence and amino acid similarity amongst some flagellar antigens of the g-complex (McQuiston et al., 2004). As well, novel gene variants were found during the study period including an *fljB* gene for antigen z53 from a subspecies IIIb isolate. While this gene variant was added to the SISTR *fljB* database prior to this study, it points to the fact that these gene databases are only as strong as the data stored in them and highlights the potential for novel gene variants to be discovered in the future. The number of incorrect calls for SISTR and SeqSero was also impacted by an issue where the *rfb* region of the genome was not fully assembled and was split over two contigs. This left the *wzx* and/or *wzy* genes that SISTR uses in its calculation absent from the assembly, leading to a null O-antigen determination, or in the case of the D1 serogroup, a B serogroup prediction. With SISTR, the B serogroup prediction was due to the carriage of the serogroup B *wzy* gene at a separate locus within the genome (Reeves et al., 2013). For SeqSero similarities in the *rfb* region outside the *wzx* lead to incorrect calls (Reeves et al., 2013). The missing sequencing information was attributable to the size selection of the genomic libraries. This region of the genome displays increased genomic fragmentation with the NexteraXT library preparation kit (data not shown), leading to insert sizes well below the 500bp minimum that was selected for. Therefore, this region was frequently filtered out of the genomic library before sequencing, leading to a gap in the sequencing data. Future implementation of *in silico* methods for *Salmonella* serotyping should ensure this size selection step is skipped in the library preparation stage to prevent bias in the sequencing library. Incorrect results from both SISTR and MLST analysis were also attributed to a lack of sufficient representative isolates from a specific serovar/and or phylogenetic lineage, throwing off the prediction. MLST analysis calls an isolate based off of the dominant serovar within its database at a specific ST. For example both Typhimurium and ssp I 4,[5],12:i:- isolates are found at ST 19. Since Typhimurium is the dominant serovar at this ST all isolates with this ST are called Typhimurium, leading to an incorrect result for ssp I 4,[5],12:i:- isolates. In SISTR similar results were noted for some ssp I 4,[5],12:b:- isolates who clustered closely with Paratyphi B var. Java representatives. Once again this issue points to the continued need to consistently update and curate the databases used to provide serovar calls.

Issues to Consider in Transition

While all the tested platforms displayed high levels of successful results, indicating their suitability in maintaining the historical records, surveillance systems, and communication structures on which *Salmonella* serotyping is based, the choice of a single *in silico* serotyping analysis platform for usage by a diagnostic and reference laboratory should be made with consideration for each method's strengths and weaknesses and

the laboratory's intended purpose. *In silico* serotyping results would be inappropriate for the investigation into the expression of antigenic factors, such as detecting all monophasics or rough isolates, but would be appropriate for routine diagnostic and reference activities. SeqSero provides a genomic understanding of the individual antigens that make up a serovar (Zhang et al., 2015), and represent the closest analogous situation to traditional serotyping. One potential advantage of SeqSero is the possibility of determining the serovar directly from the raw sequencing reads (Zhang et al., 2015). While the raw read analysis through SeqSero was not assessed in this study it may represent a positive for laboratories that lack the computational capacity to assemble genomes prior to serovar determination. MLST analysis uses ST to infer serotypes based on databases that link serovars to specific ST (Achtman et al., 2012), while this method is not analogous to traditional serotyping it allows for enhanced phylogenetic information to be generated that can be used to answer additional epidemiological questions. For example this method allows for the differentiation of *Salmonella* serovar Newport, a polyphyletic serovar, into the three distinct lineages allowing for the potential to further classify polyphyletic serovars (Achtman et al., 2012). Meanwhile SISTR utilizes both methods to provide its overall serovar determination. This not only allows users to gain an understanding of the underlying genetic carriage of the individual antigens, but also allows for the generation of enhanced phylogenetic information that can further classify the isolates (Yoshida et al., 2016b). The combination of both phylogenetic and genomic determinations of a serovar by the SISTR platform allows for significantly stronger result determination as this method led to an overall higher level of successful results as well as a reduction in the number of inconclusive matches that would require further benchtop serotyping. However, all *in silico* serotyping platforms are limited in their ability to define novel serovars, as their databases are only as expansive as what is located within them. Unique ST linkages and sequences of novel gene variants for existing or new antigens must be added to the respective databases before they can be detected and called otherwise they risk displaying an inconclusive or incorrect result.

Ultimately serotyping remains an important tool for public health officials, and the impact of integrating genomic data into existing surveillance systems cannot be ignored. In Canada, serotyping information has formed the basis on which programs such as NESP and PulseNet Canada carry out their mandates (Swaminathan et al., 2006; Government of Canada, 2014). Clinical infections of *Salmonella* are recorded at the serovar level in NESP and are compared to historical levels of disease, allowing for fast and efficient analysis of trends in disease frequencies (Government of Canada, 2014). This information can be used to quickly track *Salmonella* infections, detect outbreaks (Hutwagner et al., 1997), help determine the source of infection (Jackson et al., 2013), and give insight into disease outcomes and potential complications (Gal-Mor et al., 2014). The information provided at this level is then fed into other surveillance programs such as PulseNet Canada, providing the basis on which its databases are organized and also informs the outbreak investigations undertaken (Swaminathan et al., 2006). As Canada's national diagnostic and reference laboratories

with surveillance functions continue to move forward with the implementation of WGS to carry out their mandates, this current paradigm will be shifted and consideration must be given to the impacts this will create. WGS will provide the simultaneous collection of information that will feed into all levels of the current surveillance paradigm, negating the time factor that separates the various programs. *In silico* typing information could be collected for routine identification and diagnostics at the same time whole genome MLST or SNP results are collected for outbreak investigations allowing for the rapid fulfillment of multiple mandates. However, the loss of the time factor that is used in part to separate surveillance programs and their activities should not be used to diminish these programs mandates. Instead the separation between programs will be informed by the resolution of information generated and analyzed. Information collected by NESP on serotype disease frequencies will still be crucial in the allocation of resources spent on investigations by PulseNet Canada. As well, the organization of information within PulseNet Canada's databases will still rely on serotype information collected by NESP, and the information encoded in serotypes will still be crucial to inform questions during epidemiological investigations into outbreaks.

Further considerations must also be given to the costs associated with sequencing, the technical and informational capacities of national and partner laboratories, turnaround times associated with the batching of isolates, and data sharing models to improve the flow of information between various partners. Following completion of this study, Canada's reference laboratories are moving forward with the real-time parallel use of SISTR in comparison to traditional serotyping. Roll out of technical and informational capacities is underway within the Canadian system starting with the acquisition of Illumina MiSeq equipment at our partner laboratories and knowledge translation activities between the bioinformatics, microbiological, and epidemiological experts. Utilization of Canada's bioinformatics infrastructure and data sharing platform, IRIDA, allows stakeholders to maintain local ownership of the sample data while authorizing data sharing to specified partners. Integration of SISTR, as well as the assembly and quality control tools needed to process WGS data, into the IRIDA platform is also underway, creating a one-stop shop of data sources, analysis tools, and investigational activities.

CONCLUSION

Salmonella serotyping remains an important tool for the public health community and is integral to current public health surveillance systems in Canada. As Canada continues to transition toward using WGS to carry out its public health mandate, backward compatibility with existing surveillance systems is important. Three *in silico* serotyping platforms were validated as part of this study, SISTR, SeqSero, and MLST analysis, and we reported successful results in 94.8, 88.2, and 88.3% of the 813 isolates tested, with SISTR significantly outperforming both SeqSero and MLST. However, all platforms would be suitable for maintaining the historical records,

surveillance systems, and communication structures currently in place. Results also point to the need to reframe our understanding of serotyping within the genomic era. Use of SISTR or SeqSero provides us with an understanding of the antigenic genes that are carried by an isolate and not necessarily what is expressed by that isolate. While this may lead to incongruences between the two methods, it is important to note these incongruences are not the result of errors but just a conceptual shift in how we will be defining what a serovar is. Both SISTR and MLST also provide us with increased phylogenetic classification which can be used to answer additional epidemiological questions; while SeqSero provides the opportunity to analyze results directly from raw reads. Ultimately the choice of system will depend on users need. The adoption of WGS by diagnostic and reference laboratories with surveillance functions will provide the simultaneous collection of information that will feed into multiple levels of the current surveillance paradigm, but does not negate the importance of the various programs and the role of serotyping in these programs.

Members of the PulseNet Canada Steering Committee

The PulseNet Canada Steering Committee is made up of the following members Ana Paccagnella, Linda Hoang (BC Centre for Disease Control); Linda Chui (Alberta Provincial Laboratory for Public Health); Paul Levett, Ryan McDonald (Saskatchewan Disease Control Laboratory); John Wylie, David Alexander (Cadham Provincial Laboratory, Manitoba); Vanessa Allen, Anne Maki (Public Health Ontario); Sadjia Bekal (Laboratoire de santé publique du Québec), Ross Davidson (Queen Elizabeth II Health Sciences Centre, Nova Scotia); Elspeth Nickerson, Janet Reid (Saint John Regional Hospital, New Brunswick); Laura Gilbert (Eastern Health, Newfoundland and Labrador); Greg German (HealthPEI, Prince Edward Island); Moe Elmufi, Sean Quinlan,

REFERENCES

- Achtman, M., Wain, J., Weill, F. X., Nair, S., Zhou, Z., Sangal, V., et al. (2012). Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog.* 8:e1002776. doi: 10.1371/journal.ppat.1002776
- Ashton, P. M., Nair, S., Peters, T., Bale, J., Powell, D. G., Painset, A., et al. (2016). Identification and typing of *Salmonella* for public health surveillance using whole genome sequencing. *PeerJ* 4:e1752. doi: 10.7717/peerj.1752
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021
- Bergholz, T. M., Moreno Switt, A. I., and Wiedmann, M. (2014). Omics approaches in food safety: fulfilling the promise? *Trends Microbiol.* 22, 275–281. doi: 10.1016/j.tim.2014.01.006
- Boland, C., Bertrand, S., Mattheus, W., Dierick, K., Jasson, V., Rosseel, T., et al. (2015). Extensive genetic variability linked to IS26 insertions in the *fliB* promoter region of atypical monophasic variants of *Salmonella enterica* serovar Typhimurium. *Appl. Environ. Microbiol.* 81, 3169–3175. doi: 10.1128/aem.00270-15
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Cathy Carrillo, Ray Allain (Canadian Food Inspection Agency); Franco Pagotto (Health Canada); Lorelee Tschetter, Kim Ziebell (Public Health Agency of Canada).

AUTHOR CONTRIBUTIONS

Conceived and designed the experiment: CAY, CY, JN, ET, AR, AN, CN, and PNSC. Performed the experiments: CAY, JR, PK, and MW. Contributed analysis: CAY, CY, PK, MW, SC, and CN. Wrote the paper: CAY and CY.

FUNDING

Funding for these experiments was provided by the Public Health Agency of Canada and the Government of Canada's Genomics Research and Development Initiative.

ACKNOWLEDGMENTS

The authors would like to acknowledge: the technical and infrastructural support of Morag Graham and the Genomics Core at the NML; the bioinformatics support of Gary Van Domselaar and the Bioinformatics Core at the NML; the reference laboratories at NML Winnipeg and Guelph for isolate growth and serotyping; and Simone Gurnik at NML Guelph for help with DNA extractions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.01044/full#supplementary-material>

- Ewing, W. H. (1986). *Edwards and Ewing's Identification of Enterobacteriaceae*. New York, NY: Elsevier Science Publishing Co. Inc.
- Gal-Mor, O., Boyle, E. C., and Grassl, G. A. (2014). Same species, different diseases: how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. *Front. Microbiol.* 5:391. doi: 10.3389/fmicb.2014.00391
- Gilmour, M. W., Graham, M., Reimer, A., and Van Domselaar, G. (2013). Public health genomics and the new molecular epidemiology of bacterial pathogens. *Public Health Genomics* 16, 25–30. doi: 10.1159/000342709
- Government of Canada (2014). *National Enteric Surveillance Program (NESPP), Annual Summary 2012*. Guelph, ON: Public Health Agency of Canada.
- Grimont, P. A. D., and Weill, F.-X. (2007). "Antigenic formulae of the *Salmonella* serovars," in *Proceedings of the WHO Collaborating Center for Reference and Research on Salmonella*, 9th Edn, (Paris: Institut Pasteur).
- Hendriksen, R. S., Mikoleit, M., Carlson, V. P., Karlsmose, S., Vieira, A. R., Jensen, A. B., et al. (2009). WHO Global Salm-Surv external quality assurance system for serotyping of *Salmonella* isolates from 2000 to 2007. *J. Clin. Microbiol.* 47, 2729–2736. doi: 10.1128/jcm.02437-08
- Hendriksen, R. S., Vieira, A. R., Karlsmose, S., Lo Fo Wong, D. M., Jensen, A. B., Wegener, H. C., et al. (2011). Global monitoring of *Salmonella* serovar distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: results of quality assured laboratories from 2001 to 2007. *Foodborne Pathog. Dis.* 8, 887–900. doi: 10.1089/fpd.2010.0787

- Hutwagner, L. C., Maloney, E. K., Bean, N. H., Slutsker, L., and Martin, S. M. (1997). Using laboratory-based surveillance data for prevention: an algorithm for detecting *Salmonella* outbreaks. *Emerg. Infect. Dis.* 3, 395–400. doi: 10.3201/eid0303.970322
- Jackson, B. R., Griffin, P. M., Cole, D., Walsh, K. A., and Chai, S. J. (2013). Outbreak-associated *Salmonella enterica* serotypes and food Commodities, United States, 1998–2008. *Emerg. Infect. Dis.* 19, 1239–1244. doi: 10.3201/eid1908.121511
- Kim, S., Frye, J. G., Hu, J., Fedorka-Cray, P. J., Gautam, R., and Boyle, D. S. (2006). Multiplex PCR-based method for identification of common clinical serotypes of *Salmonella enterica* subspecies enterica. *J. Clin. Microbiol.* 44, 3608–3615. doi: 10.1128/JCM.00701-06
- Kong, Q., Yang, J., Liu, Q., Alamuri, P., Roland, K. L., and Curtiss, R. (2011). Effect of deletion of genes involved in lipopolysaccharide core and O-antigen synthesis on virulence and immunogenicity of *Salmonella enterica* serovar Typhimurium. *Infect. Immun.* 79, 4227–4239. doi: 10.1128/iai.05398-11
- Mackinnon, A. (2000). A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. *Comput. Biol. Med.* 30, 127–134. doi: 10.1016/S0010-4825(00)00006-8
- Magoc, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507
- Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3140–3145. doi: 10.1073/pnas.95.6.3140
- McQuiston, J. R., Parrenas, R., Ortiz-Rivera, M., Gheesling, L., Brenner, F., and Fields, P. I. (2004). Sequencing and comparative analysis of flagellin genes *fliC*, *fliB*, and *fliP* from *Salmonella*. *J. Clin. Microbiol.* 42, 1923–1932. doi: 10.1128/JCM.42.5.1923-1932.2004
- Mikoleit, M., Van Duyne, M. S., Halpin, J., Mcglinchey, B., and Fields, P. I. (2012). Variable expression of O:61 in *Salmonella* group C2. *J. Clin. Microbiol.* 50, 4098–4099. doi: 10.1128/jcm.01676-12
- Parmley, E. J., Pintar, K., Majowicz, S., Avery, B., Cook, A., Jokinen, C., et al. (2013). A Canadian application of one health: integration of *Salmonella* data from various Canadian surveillance programs (2005–2010). *Foodborne Pathog. Dis.* 10, 747–756. doi: 10.1089/fpd.2012.1438
- Reeves, P. R., Cunneen, M. M., Liu, B., and Wang, L. (2013). Genetics and evolution of the *Salmonella* galactose-initiated set of O antigens. *PLoS ONE* 8:e69306. doi: 10.1371/journal.pone.0069306
- Reyes, R. E., Gonzalez, C. R., Jimenez, R. C., Herrera, M. O., and Andrade, A. A. (2012). “Mechanism of O-antigen structural variation of bacterial lipopolysaccharide (LPS)” in *The Complex World of Polysaccharides*, ed. D. N. Karunaratne (Rijeka: InTech), doi: 10.5772/48147
- Ronholm, J., Nasheri, N., Petronella, N., and Pagotto, F. (2016). Navigating microbiological food safety in the era of whole-genome sequencing. *Clin. Microbiol. Rev.* 29, 837–857. doi: 10.1128/cmr.00056-16
- Shipp, C. R., and Rowe, B. (1980). A mechanised microtechnique for *Salmonella* serotyping. *J. Clin. Pathol.* 33, 595–597. doi: 10.1136/jcp.33.6.595
- Swaminathan, B., Gerner-Smith, P., Ng, L. K., Lukinmaa, S., Kam, K. M., Rolando, S., et al. (2006). Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. *Foodborne Pathog. Dis.* 3, 36–50. doi: 10.1089/fpd.2006.3.36
- Thomas, M. K., Murray, R., Flockhart, L., Pintar, K., Fazil, A., Nesbitt, A., et al. (2015). Estimates of foodborne illness-related hospitalizations and deaths in Canada for 30 specified pathogens and unspecified agents. *Foodborne Pathog. Dis.* 12, 820–827. doi: 10.1089/fpd.2015.1966
- Thomas, M. K., Murray, R., Flockhart, L., Pintar, K., Pollari, F., Fazil, A., et al. (2013). Estimates of the burden of foodborne illness in Canada for 30 specified pathogens and unspecified agents, circa 2006. *Foodborne Pathog. Dis.* 10, 639–648. doi: 10.1089/fpd.2012.1389
- Toboldt, A., Tietze, E., Helmuth, R., Junker, E., Fruth, A., and Malorny, B. (2013). Population structure of *Salmonella enterica* serovar 4,[5],12:b:- strains and likely sources of human infection. *Appl. Environ. Microbiol.* 79, 5121–5129. doi: 10.1128/aem.01735-13
- Wattiau, P., Boland, C., and Bertrand, S. (2011). Methodologies for *Salmonella enterica* subsp. *enterica* subtyping: gold standards and alternatives. *Appl. Environ. Microbiol.* 77, 7877–7885. doi: 10.1128/aem.05527-11
- World Health Organization (2015). *WHO Estimates of the Global Burden of Foodborne Diseases: Foodborne Disease Burden Epidemiology Reference Group 2007–2015*. Geneva: World Health Organization.
- Yoshida, C., Gurnik, S., Ahmad, A., Blimkie, T., Murphy, S. A., Kropinski, A. M., et al. (2016a). Evaluation of molecular methods for the identification of *Salmonella* serovars. *J. Clin. Microbiol.* 54, 1992–1998. doi: 10.1128/jcm.00262-16
- Yoshida, C., Kruczakiewicz, P., Laing, C. R., Lingohr, E. J., Gannon, V. P. J., Nash, J. H. E., et al. (2016b). The *Salmonella* in silico typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS ONE* 11:e0147101. doi: 10.1371/journal.pone.0147101
- Yoshida, C., Lingohr, E. J., Trognitz, F., Maclarens, N., Rosano, A., Murphy, S. A., et al. (2014). Multi-laboratory evaluation of the rapid genosubtyping array (SGSA) for the identification of *Salmonella* serovars. *Diagn. Microbiol. Infect. Dis.* 80, 185–190. doi: 10.1016/j.diagmicrobio.2014.08.006
- Zhang, S., Yin, Y., Jones, M. B., Zhang, Z., Deatherage Kaiser, B. L., Dinsmore, B. A., et al. (2015). *Salmonella* serotype determination utilizing high-throughput genome sequencing data. *J. Clin. Microbiol.* 53, 1685–1692. doi: 10.1128/jcm.00323-15

Conflict of Interest Statement: CY, JN, PK, and ET were part of the SISTR development team.

The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Yachison, Yoshida, Robertson, Nash, Kruczakiewicz, Taboada, Walker, Reimer, Christianson, Nichani, The PulseNet Canada Steering Committee and Nadon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Pan-genome Analyses of the Species *Salmonella enterica*, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar

Chad R. Laing*, Matthew D. Whiteside and Victor P. J. Gannon

National Microbiology Laboratory, Public Health Agency of Canada, Lethbridge, AB, Canada

OPEN ACCESS

Edited by:

Sandra Torriani,
University of Verona, Italy

Reviewed by:

Jinshui Zheng,
Huazhong Agricultural University,
China
Dapeng Wang,
Shanghai Jiao Tong University, China

***Correspondence:**

Chad R. Laing
chadr.laing@canada.ca

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 20 February 2017

Accepted: 03 July 2017

Published: 31 July 2017

Citation:

Laing CR, Whiteside MD and
Gannon VPJ (2017) Pan-genome
Analyses of the Species *Salmonella*
enterica, and Identification
of Genomic Markers Predictive
for Species, Subspecies,
and Serovar. *Front. Microbiol.* 8:1345.
doi: 10.3389/fmicb.2017.01345

Food safety is a global concern, with upward of 2.2 million deaths due to enteric disease every year. Current whole-genome sequencing platforms allow routine sequencing of enteric pathogens for surveillance, and during outbreaks; however, a remaining challenge is the identification of genomic markers that are predictive of strain groups that pose the most significant health threats to humans, or that can persist in specific environments. We have previously developed the software program Panseq, which identifies the pan-genome among a group of sequences, and the SuperPhy platform, which utilizes this pan-genome information to identify biomarkers that are predictive of groups of bacterial strains. In this study, we examined the pan-genome of 4893 genomes of *Salmonella enterica*, an enteric pathogen responsible for the loss of more disability adjusted life years than any other enteric pathogen. We identified a pan-genome of 25.3 Mbp, a strict core of 1.5 Mbp present in all genomes, and a conserved core of 3.2 Mbp found in at least 96% of these genomes. We also identified 404 genomic regions of 1000 bp that were specific to the species *S. enterica*. These species-specific regions were found to encode mostly hypothetical proteins, effectors, and other proteins related to virulence. For each of the six *S. enterica* subspecies, markers unique to each were identified. No serovar had pan-genome regions that were present in all of its genomes and absent in all other serovars; however, each serovar did have genomic regions that were universally present among all constituent members, and statistically predictive of the serovar. The phylogeny based on SNPs within the conserved core genome was found to be highly concordant to that produced by a phylogeny using the presence/absence of 1000 bp regions of the entire pan-genome. Future studies could use these predictive regions as components of a vaccine to prevent salmonellosis, as well as in simple and rapid diagnostic tests for both *in silico* and wet-lab applications, with uses ranging from food safety to public health. Lastly, the tools and methods described in this study could be applied as a pan-genomics framework to other population genomic studies seeking to identify markers for other bacterial species and their sub-groups.

Keywords: genomics, pan-genome, *Salmonella*, predictive markers, food safety

INTRODUCTION

The global burden of bacterial enteric disease, much of it foodborne, results in an estimated 2.2 million deaths per year, and an annual loss of 112,000 disability adjusted life years in the United States alone (Bergholz et al., 2014; Scallan et al., 2015). Nationwide molecular diagnostic networks, such as PulseNet in North America, were designed to enable the rapid identification of outbreaks by genetic fingerprinting the etiological agents of disease, and keeping nationwide databases of genetic fingerprints of specific pathogens associated with human disease. Since its inception, PulseNet has relied on pulsed-field gel electrophoresis (PFGE) for fingerprinting of bacterial pathogens to identify the specific sources of outbreaks and prevent further infections. Using this approach, it has been estimated that PulseNet prevents 277,000 illnesses from bacterial pathogens annually in the United States, reducing the costs associated with medical care and loss of productivity due to worker illness (Scharff et al., 2016).

Despite the usefulness of PulseNet, the PFGE technique itself is often unable to distinguish between related and unrelated strains, due to its reliance on rare-cutting restriction enzyme sites within the genome (Allard et al., 2012). Additionally, the interpretation of the banding patterns among labs requires extensive training and standardization to enable meaningful comparisons. Lastly, the banding patterns provide no information on the actual content of the genomes they represent, so important information regarding human virulence, such as the presence or absence of known toxins, is not available.

Lastly, while the presence of known virulence factors has been correlated with severe human disease in a number of bacterial species, it has also been shown that some lineages or clades within these same species, while possessing specific virulence factors, are rarely associated with human disease (Lupolova et al., 2016; Waryah et al., 2016). Thus, multiple virulence factors, and regulatory genes that influence the expression of key virulence factors, or otherwise modulate the virulence of these strains, need to be taken into consideration when attempting to predict the strains of a bacterial species that are potential human health threats (Oprijnen et al., 2012).

Recently, whole-genome sequencing (WGS) has displaced PFGE as the *de facto* standard for the complete characterization of bacterial pathogens, in both ongoing surveillance and outbreak investigations (Deng et al., 2016; Franz et al., 2016). WGS allows clear definition between outbreak-related strains and those from unrelated sources, and it has the ability to identify routes of transmission, and attribute bacterial contaminants to specific sources (den Bakker et al., 2014). It is currently being utilized in reference laboratories worldwide. Examples of its application include the sequencing of all *Listeria monocytogenes* isolated in the United States, all *Salmonella* isolated by the Food and Drug Administration in the USA, and by Public Health England as part of routine surveillance (Ashton et al., 2016), and a large-scale survey of *Staphylococcus aureus* in continental Europe. In the latter study, the applicability of WGS for the identification of the emergence and spread of clinically relevant *Staphylococcus aureus* was demonstrated (Aanensen et al., 2016).

It has also recently been shown that antimicrobial resistance (Tyson et al., 2015; McDermott et al., 2016; Zhao et al., 2016), serovar (Levine et al., 2016; Yoshida et al., 2016b), and the results of other traditional sub-typing schemes such as multi-locus sequence typing (Sheppard et al., 2012) can be accurately predicted *in silico* through the analysis of bacterial genome sequences. However, identifying bacterial isolates that are most likely to cause disease in humans, based on the genome sequence alone, is a more complex task. In addition, markers that can identify bacteria likely to exhibit particular phenotypes, such as the ability to survive in a particular niche, or the ability to tolerate harsh environments such as those found in food processing plants are also required.

We have previously developed the software platform Panseq, for the analyses of thousands of genomes in a pan-genome context, where both the presence/absence of the accessory genome and SNPs within the shared core-genome are computed (Laing et al., 2010). Additionally, we recently released a platform for the predictive genomics of *Escherichia coli*, called SuperPhy, in which markers statistically biased within groups of bacteria, based on any metadata category, can be identified (Whiteside et al., 2016).

In this study we use our previously created software to examine the pan-genome of *Salmonella enterica*, a pathogen that causes an estimated 93.8 million cases of enteric illness worldwide each year (Majowicz et al., 2010; Gal-Mor et al., 2014). The species *S. enterica* is divided into six subspecies: *enterica*, *salamae*, *arizonae*, *diarizonae*, *houtenae*, and *indica*. Over 99% of human disease caused by *S. enterica* is done so by subspecies *enterica*, with the World Health Organization estimating that *S. enterica* infections from contaminated food alone constitute a loss of 6.43 million disability adjusted life years worldwide, more than any other enteric pathogen (Kirk et al., 2015). Within this bacterial subspecies, are human-adapted strains responsible for typhoid fever, as well as a large number of animal-derived non-typhoidal strains responsible for foodborne illness. In this study, we have identified species- and subspecies-specific markers, as well as markers predictive of serovar for subspecies *enterica*. While this study focused on *S. enterica*, the tools and approach are broadly applicable to any species or collection of genomes.

MATERIALS AND METHODS

All commands and parameters used to analyze the data and generate the Figures are available as Supplementary File 1. The scripts used for analyses are available at <https://github.com/superphy/gamechanger>. The following is a summary of the methods used.

Data Collection

All *S. enterica* genomes were downloaded from GenBank in nucleotide fasta format. A full listing of the initial 4939 genomes, including GenBank identifier, subspecies, serovar, the number of species-specific core regions present, the number of contigs, and whether the genome passed the quality filtering steps are listed in Supplementary File 2.

Serovar Identification

Most of the *S. enterica* genomes in GenBank had serovar provided as part of their metadata; however, 321 were missing this designation. The SISTR web-server, as well as the SISTR commandline app were used to predict the serovar for these strains (Yoshida et al., 2016b).

Pan-genome Analyses

Panseq (commit:1d0ab9d37e8e358d266e1d0aa80e9b27f28a1def) was used to identify the pan-genome of the 4939 strains in this study (Laing et al., 2010). Genomes were initially fragmented into 1000 bp segments, and subsequently clustered using cd-hit v.4.6 to remove potential duplicates/paralogs from the analyses using a 90% sequence identity threshold (Fu et al., 2012). Initially Panseq was used to determine the distribution of the pan-genome among the genomes at a 90% sequence identity threshold, from which a “conserved core” was identified. Within the conserved core, Panseq was then used to identify single-nucleotide polymorphisms.

Identification of *S. enterica* Species-Specific Regions

To identify regions that were likely to represent the species as a whole, we initially examined the 211 closed *S. enterica* genomes in GenBank (Supplementary File 2), and identified 3832 regions of 1000 bp that were found in 90% (190) of the 211 closed genomes using Panseq, at a 90% sequence identity threshold. These regions were then screened against the online GenBank nr database using megablast as a first-pass filter with default parameters, searching across bacteria (taxid:2), and excluding all *Salmonella* (taxid:590) hits that had greater than 80% identity across 80% of the query length from the results. The remaining 1482 genomic regions were subsequently screened against the online GenBank nr database of all bacteria (taxid:2), using the blastn algorithm, to identify matches that were missed using the less-specific megablast algorithm, with word size 11, an e-value cutoff of 0.001, and excluding all *Salmonella* (taxid:590). These results were filtered in the same manner, leaving 405 potentially species-specific regions. Lastly, these regions were compared against *Salmonella bongori* genomes in GenBank; one *S. bongori* hit was identified, which left 404 genomic regions present in *S. enterica* but no other bacterial genomic sequences within the GenBank nr database.

The putative function of these regions was determined by screening them across the GenBank nr database using blastx with “max hits:10,” “taxid limit:1236 (gammaproteobacteria),” and an “e-value threshold: 0.001.” The best matching hit above a 90% sequence identity threshold was used for the putative functional assignment.

Identification of Subspecies- and Serovar-Specific Regions

The Fisher’s Exact test, using the Bonferroni correction for multiple testing was applied as in the SuperPhy platform (Whiteside et al., 2016), implemented here as the standalone

program feht¹. The input for the program was Supplementary File 2, which contained metadata for all the strains, as well as the binary_table.txt output file from the Panseq analyses, which denotes the presence/absence of each 1000 bp pan-genome region among all the strains.

S. enterica Phylogenetic Analyses

The phylogeny based on SNPs within the core genome was generated using RAxML v8.2.9, with the.snp.phyip output file from Panseq (Stamatakis, 2014). The phylogeny based on the presence/absence of the pan-genome was also generated using RAxML v8.2.9, with the binary.phyip output file from Panseq.

Generation of Figures and Tables

The R-statistical language v3.3.2 was used to generate the summary Figures and Tables (R Core Team, 2016). The R-scripts and all others used for the analyses can be found at <https://github.com/superphy/gamechanger/tree/master/src>. The ggtree package for R was used in the generation of the phylogenetic tree images (Yu et al., 2016).

RESULTS

S. enterica Pan-genome

We initially determined the size and distribution of the *S. enterica* pan-genome as genome fragments of 1000 bp in size, across the 4939 genome sequences of this study, which are summarized by subspecies in **Table 1**, and within subspecies *enterica* by serovar in **Table 2**. As can be seen in **Figure 1**, the pan-genome comprised of 4939 *S. enterica* genomes was found to be 25.3 Mbp in size, with 70% of the pan-genome present in fewer than 100 strains. Conversely, the core genome was found to be 1.5 Mbp in size, with all but 200 genomes (96%) containing 3.2 Mbp of shared genomic core. Only 17% of the pan-genome was found in greater than 100 genomes, but fewer than 4739 genomes.

S. enterica Species-Specific Regions

To identify regions of *S. enterica* that were likely to be shared among most genomes of the species, we examined all 211 closed genomes of *S. enterica* in GenBank, looking for genomic regions that were present in at least 190 (90%) of these genomes. We

¹<https://github.com/chadlaing/feht>

TABLE 1 | The frequency of the subspecies observed within the study set of 4936 *Salmonella enterica* genomes, prior to any quality filtering.

Subspecies	No.
<i>enterica</i>	4913
<i>arizonae</i>	7
<i>dairizonae</i>	7
<i>houtenae</i>	4
<i>salamae</i>	4
<i>indica</i>	1

TABLE 2 | The serovars with more than 20 representatives in the current study set of 4936 *Salmonella enterica* genomes, and their frequency, prior to any quality filtering.

Serovar	No.
Typhi	1977
Typhimurium	758
Enteritidis	413
Heidelberg	201
Paratyphi	158
Kentucky	155
Agona	136
Weltevreden	120
Bareilly	106
Newport	82
Tennessee	77
Montevideo	69
Saintpaul	48
Infantis	39
Senftenberg	35
Bovismorbificans	34
Hadar	33
Muenchen	30
Anatum	27
Schwarzengrund	27
Dublin	24
Cerro	21

The list of all serovars and their frequency within the current study is available as Supplementary File 2.

identified 3832 regions of 1000 bp that were present in at least 90% of the closed genomes. These regions were subsequently screened against the GenBank nr database, and any present in non-*Salmonella* genomes were removed, leaving 404 putative *S. enterica* species-specific regions (Supplementary File 3).

Figure 2 shows the carriage of these 404 regions among the 4939 genomes of this study. All but 105 genomes contained at least 330 of these putative *S. enterica* specific regions. A stark difference in carriage of these species-specific markers was observed, with 4742 genomes containing at least 350 species-specific markers, while only 2674 genomes contained 360 or more species-specific markers.

Quality Filtering for Subsequent Analyses

To ensure the quality of the genomes in use for subsequent analyses, we plotted carriage of the 404 species-specific regions versus the number of contigs that each sequenced genome was comprised of (**Figure 3**). As can be seen, the two genomes marked in yellow contained only one, and the same, species-specific region each, despite being comprised of relatively few contigs. Subsequent searches against the GenBank nr database identified these two genomes as *Citrobacter* spp. contamination, mislabeled as *S. enterica* (GCA_001570325 and GCA_001570345). The “*Salmonella enterica* species-specific region” found in both of the contaminant *Citrobacter* genomes, did not match any other *Citrobacter* spp. in GenBank above the thresholds used for determining presence/absence in this study. However, due to the

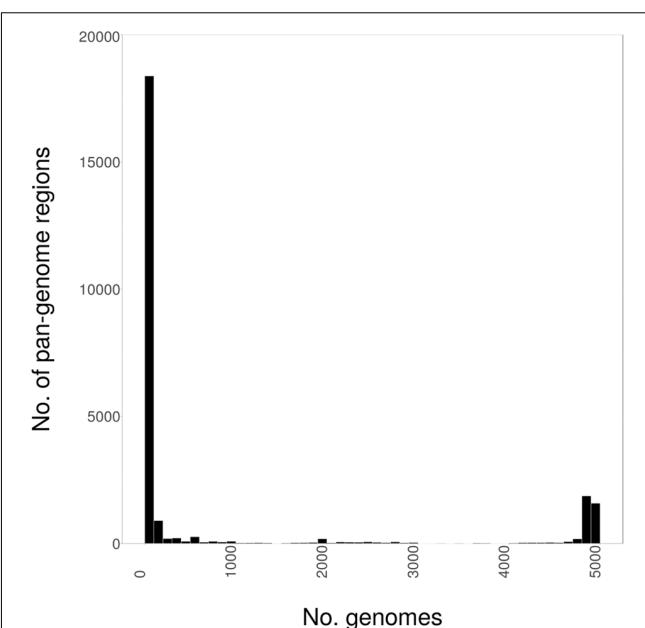


FIGURE 1 | The distribution of the *Salmonella enterica* pan-genome, as 1000 bp fragments, among 4939 whole-genome sequences (WGSs).

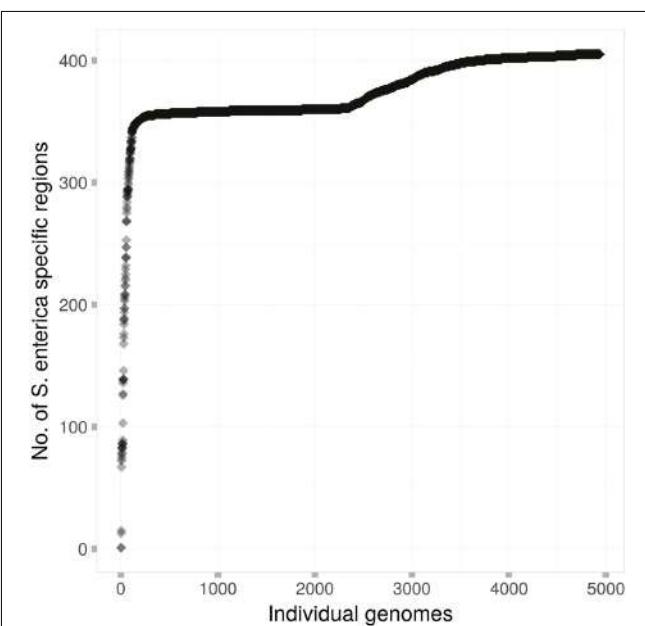
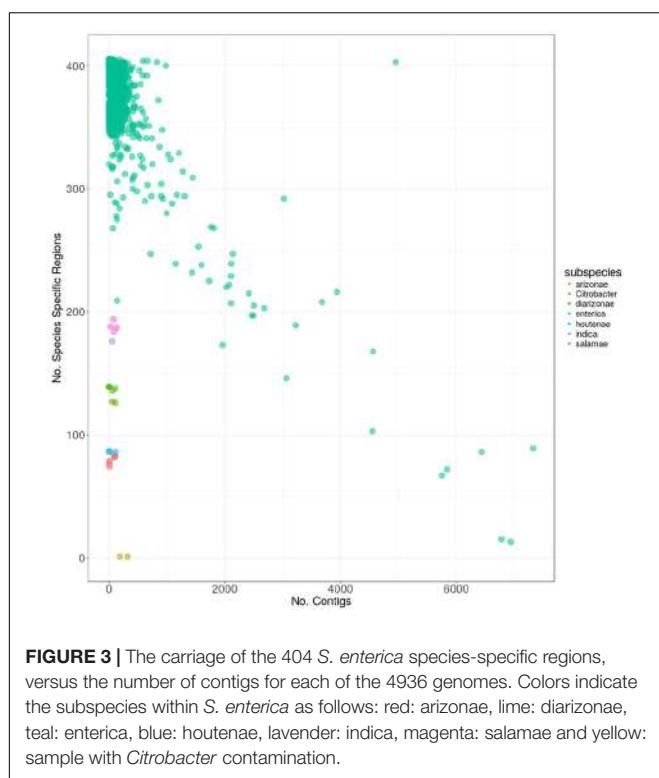


FIGURE 2 | The carriage of the 404 *S. enterica* species-specific regions among each of the 4939 genomes of this study. Each dot represents a single *S. enterica* genome, which are arranged in order from those that contain the fewest species-specific regions to those that contain the most.

presence of this region in what have been identified as *Citrobacter* genomes, the region was removed from subsequent analyses.

The majority of genomes (4913) were from subspecies *enterica*, with genomes from the five other *S. enterica* subspecies present in drastically fewer numbers (**Table 1**). All closed



genomes from subspecies *enterica* contained greater than 250 species-specific regions, which was more than the genomes from any other subspecies, with the exception of *enterica* genomes that were of poor quality and comprised of many 1000s of contigs (**Figure 3**). Genomes from subspecies *houtenae* and *arizonae* contained fewer than 100 species-specific regions, while genomes from *diarizonae*, *indica*, and *salamae* contained between 100 and 200 species-specific regions. All regions were screened against *S. bongori* to ensure specificity to *S. enterica*; one region was found to also be present in genomes from *S. bongori* and was removed from further analyses.

Within subspecies *enterica*, a negative linear relationship was observed among the number of species-specific regions contained within a genome, and the number of contigs the genome was comprised of, with the worst-case genome (GCA_000495155) being comprised of 6945 contigs, but containing only 13 species-specific regions. Other genomes such as *S. enterica* Bovismorbificans strain GCA_001114865 contained both few contigs (140) as well as fewer species-specific regions (209) than other *enterica* genomes. Additional searches discovered sequencing gaps within the genome totaling over 464 Kbp. A final outlier genome harbored nearly 5000 contigs, but also contained 403 of the species-specific regions. It was determined by searching the GenBank database, that this sequence (GCA_000765055) was actually a combination of multiple genomes in a single file.

Given the above information, all genomes from the five subspecies other than *enterica* were included in subsequent analyses, while the thresholds for inclusion of *enterica* genomes were set at a maximum of 1000 contigs, and a minimum of

250 species-specific regions. Following this quality filtering, 43 genomes were removed, leaving 4870 *S. enterica* subspecies *enterica* genomes for the following analyses.

Phylogeny of *S. enterica* Using the Conserved Core Genome

Based on the distribution of the pan-genome presented in **Figure 1**, the “conserved core” of *S. enterica* was set at being present in more than 4500 genomes, to fully capture the conserved genomic regions within the species. A phylogeny based on the SNPs among these shared regions was created, and is shown along with the distribution of the *S. enterica* species-specific regions in **Figure 4**. As can be seen, the majority of the genomes are subspecies *enterica*, and the other five subspecies are relatively more distant in the order of *indica*, *salamae*, *houtenae*, *diarizonae*, and *arizonae*. However, the order of subspecies in declining number of species-specific regions is: *enterica*, *diarizonae*, *salamae*, *indica*, *houtenae*, and *arizonae*, which is shown in **Figure 3**.

The serovar distribution within subspecies *enterica* was shown to be largely concordant with phylogeny, as demonstrated in **Figure 5**, where the 10 most abundant serovars in the current study are highlighted. However, not all serovars clustered as monophyletic groups, as can be seen with serovar Bareilly; nor were all clades found to be comprised of single serovars, demonstrated by the clade containing genomes of serovars Bareilly and Agona.

The large clades within the phylogenetic tree also demonstrate clade-specific patterns of presence/absence for the 404 species-specific markers. Among the most abundant serovars, Typhimurium, Heidelberg, Newport, and Enteritidis were found to contain the most species-specific markers, and grouped together near the center of the tree. Likewise, serovars Agona, Welevreden, and Kentucky contained fewer species-specific regions, and group together near the bottom of the tree, closer to the non-*enterica* sub-species genomes.

Table 3 considers all serovars with at least 10 members in the dataset, and the average number of species-specific markers per serovar. As can be seen, the serovars with the largest average number of species-specific regions were: Enteritidis (401.7), Anatum (401.5), Muenchen (400.5), Hadar (400.3), and Typhimurium (400.1); conversely, the serovars with the fewest average number of species-specific regions were: Derby (360.7), Montevideo (360.1), Typhi (358.1), Bovismorbificans (355.3), and Cerro (342.0).

Phylogeny of *S. enterica* Using the Pan-genome

A phylogeny based on the presence/absence of the pan-genome among the 4893 *S. enterica* genomes was created, and is shown along with the distribution of the *S. enterica* species-specific regions in **Figure 6**. As can be seen this phylogeny based on the presence/absence of the entire 25.3 Mbp pan-genome is highly concordant with the phylogeny based on the SNPs found in the conserved core of the same strains (**Figure 5**). In both trees the serovars cluster together and in the same relation to each

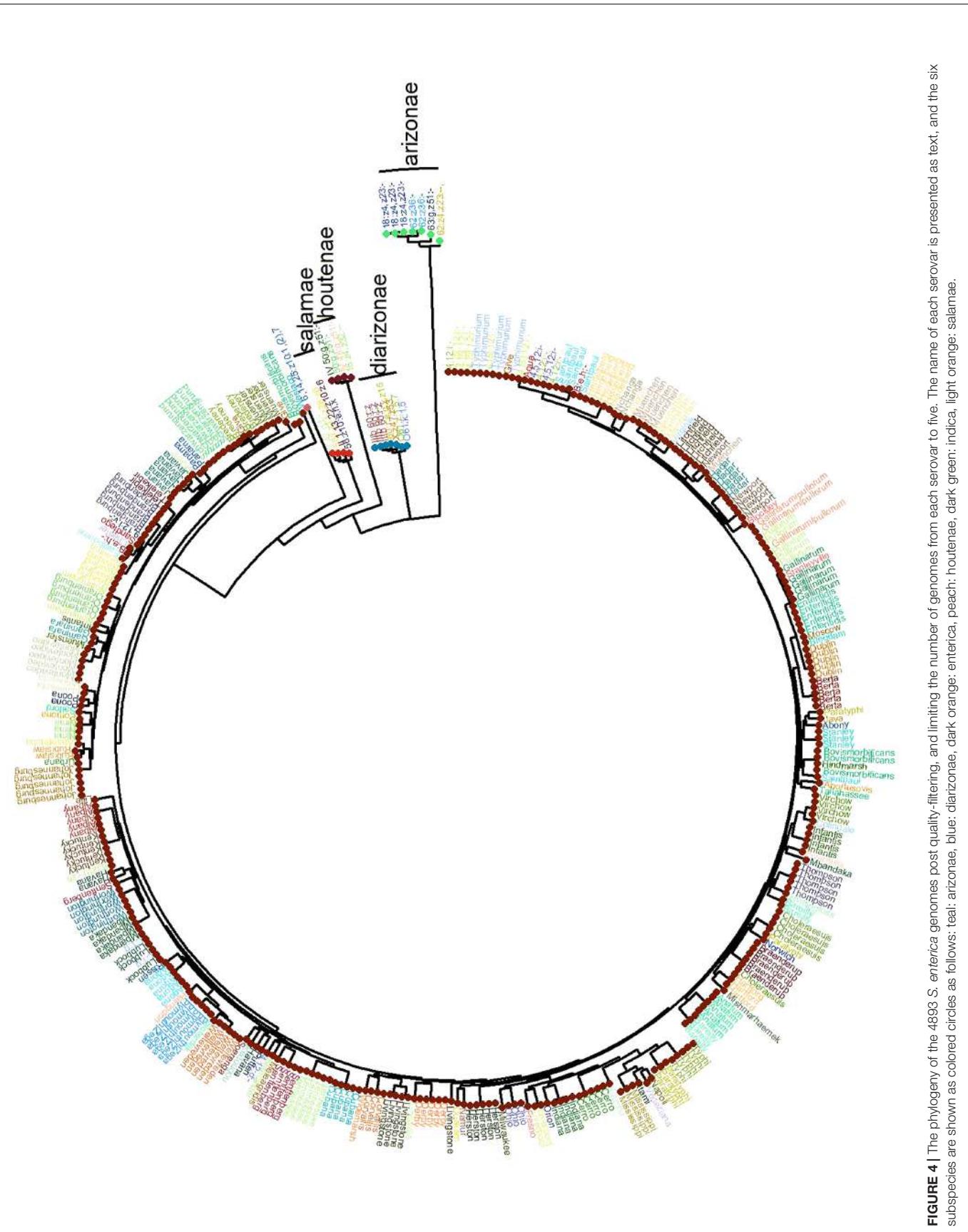


FIGURE 4 | The phylogeny of the 4893 *S. enterica* genomes post quality-filtering, and limiting the number of genomes from each serovar to five. The name of each serovar is presented as text, and the six subspecies are shown as colored circles as follows: teal: arizonae; light green: houtae; dark green: salamae; peach: diarizoneae; dark orange: enterica; blue: all others.

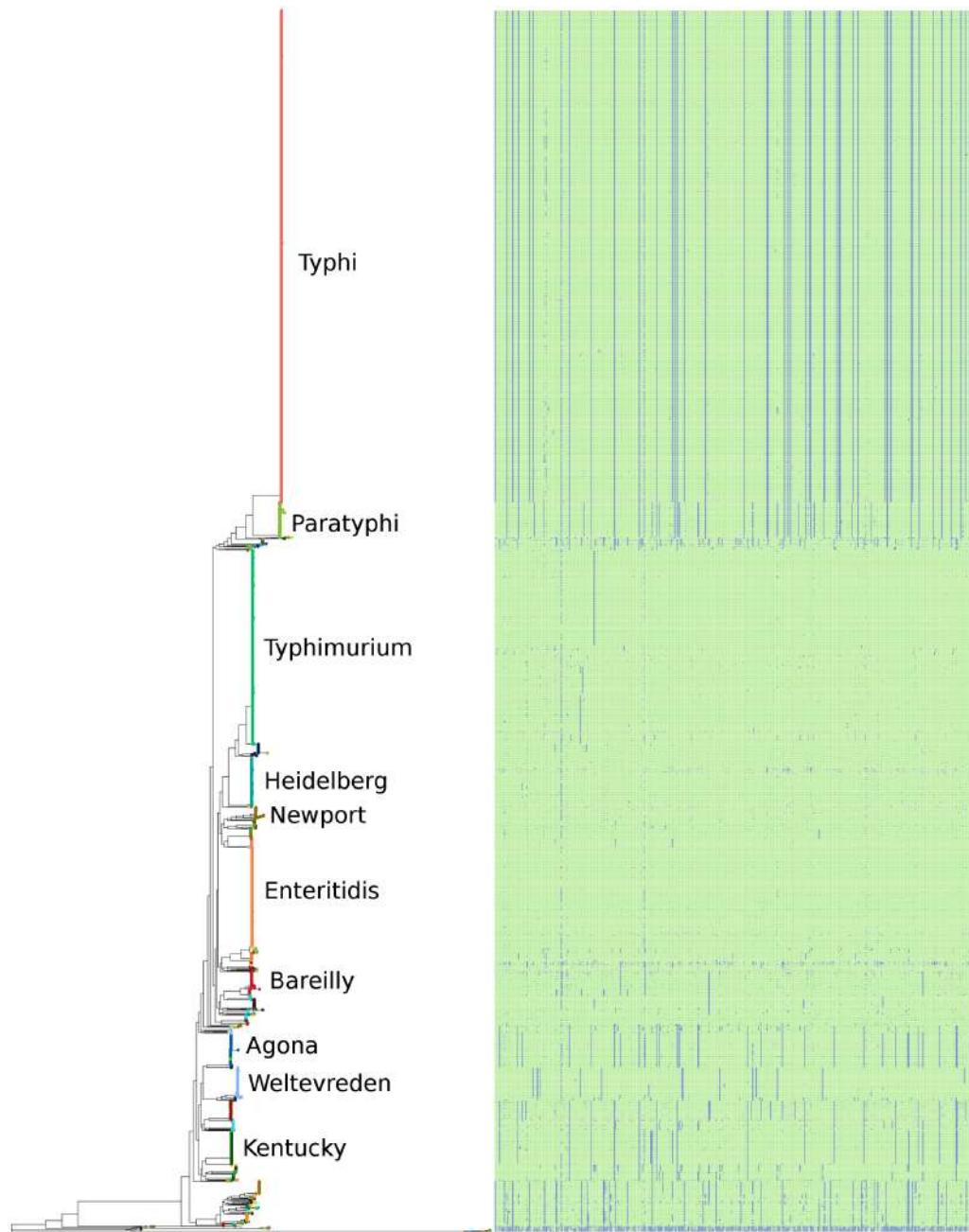


FIGURE 5 | The phylogeny of the 4893 *S. enterica* genomes post quality-filtering based on SNPs found within the conserved core genome. The 10 most abundant serovars of subspecies enterica in the current study (Agona, Bareilly, Enteritidis, Heidelberg, Kentucky, Newport, Paratyphi, Typhi, Typhimurium, Weltevreden) are labeled on the tree. The matrix to the right of the phylogeny represents the 404 species-specific regions, with blue being the absence of a region, and green being the presence of a region, for each of the genomes of the study.

other, for example serovars Typhi and Paratyphi strains form a discrete monophyletic clade. However, the branch lengths in the pan-genome tree are larger than those in the conserved SNP tree, due to the larger variation among the presence/absence of the pan-genome than to sequence variation among shared core regions.

Identification of a Minimum Set of Species-Specific Genomic Markers

Within the 404 species-specific markers, none were specific for any of the subspecies. That is, a marker was always present in genomes from at least two subspecies.

TABLE 3 | The average number of species-specific genomic regions found among serovars of subspecies *enterica*, that contained at least 10 representative genomes, within the 4870 quality filtered subspecies *enterica* genomes of this study.

Serovar	Average no. species-specific regions
Enteritidis	401.7
Anatum	401.5
Muenchen	400.5
Hadar	400.3
Typhimurium	400.1
Newport	399.8
Thompson	399.7
Saintpaul	399.6
Heidelberg	397.4
Dublin	395.2
Infantis	394.9
Braenderup	392.8
Weltevreden	390.0
Bareilly	388.5
Kentucky	380.3
Plymouth/Zega	377.9
Senftenberg	376.5
Mbandaka	374.5
Lubbock	374.1
Reading	370.4
Agona	369.5
Tennessee	368.3
Schwarzengrund	362.3
Paratyphi	361.5
Derby	360.7
Montevideo	360.1
Typhi	358.1
Bovismorbificans	355.3
Cerro	342.0

We next determined that the presence of a minimum set of two genomic regions was required to unambiguously identify genomes of *S. enterica*, within the 4893 genomes of the current study. A combination of two genomic regions were all that was required, and two such markers that were also present in the most *S. enterica* genomes were found at the following locations within the Typhimurium reference genome LT2: (1336001.. 1337000) and (2467001.. 2468000) (Supplementary File 3). All members of *S. enterica* examined contained at least one of these markers, but many other combinations within the 404 species-specific markers are also possible.

Putative Functional Identification of the *S. enterica* Species-Specific Regions

The putative function of the 404 quality-filtered *S. enterica* species-specific regions were determined from the GenBank nr database. The annotation of each of the 404 regions is available

as Supplementary File 1. **Table 4** summarizes the frequency of functional annotation categories, after annotating each region with the single best match. As can be seen, hypothetical proteins accounted for the majority (64) of the 404 annotations, with secreted effector and membrane proteins being the next most frequent category among the species-specific regions. Other membrane, transport, and secretion proteins were observed. The species-specific regions also included proteins involved in core metabolic functions, protein and DNA synthesis, and response to stress.

Identification of Subspecies-Specific Markers from the Pan-genome

Having identified species-specific markers, we employed the same techniques, utilizing the presence/absence of all pan-genome markers, just as was carried out in identifying the 404 species-specific ones, to identify subspecies-specific markers. The number of markers that were completely unique to a subspecies is given in **Table 5**. Subspecies *arizonaee* contained the most unique markers, at 207, and *enterica* contained the least, at 9.

Identification of Universal Serovar Markers within Subspecies *enterica* from the Pan-genome

Subspecies *enterica* genomes were the vast majority of those available, so we attempted to identify serovar-specific markers for the top 10 serovars, in the same manner that we identified subspecies-specific markers. We found that there were no genomic markers that uniquely defined any of the serovars based on their presence or absence; however, there were a number of genomic regions that were universally present or absent among serovars, as well as statistically over- or under-represented with respect to all other serovar genomes from this study; they are shown in **Table 6**.

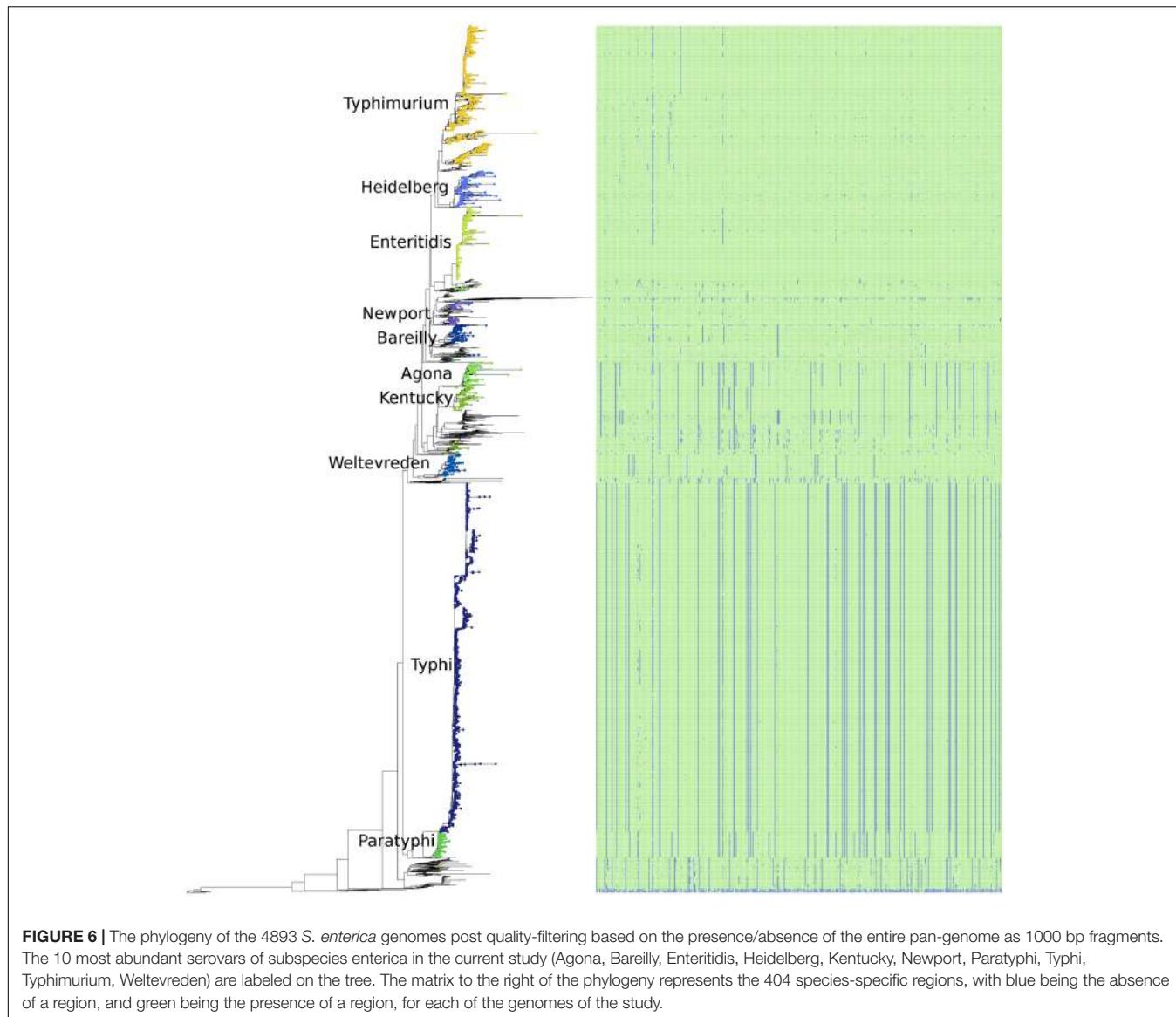
To further assess the validity of these markers, a dataset comprised of 3948 genomes from Enterobase², was selected to have an identical number of strains belonging to each of nine serovars in our GenBank dataset. The Enterobase dataset was used to test the predictive markers we identified from the GenBank dataset in the first part of the study. The results of this comparison are shown in **Figure 7**. As can be seen, the markers were well-conserved among the Enterobase dataset, with eight of the nine serovars having a subset of the predictive markers present among all of the test genomes; serovar Typhimurium had a marker subset that was present in all but one of the test genomes.

DISCUSSION

S. enterica Pan-genome

Previous examinations of the *S. enterica* pan-genome were based on relatively small datasets of 45 and 73 genomes (Jacobsen

²<https://enterobase.warwick.ac.uk/species/index/senterica>



et al., 2011; Leekitcharoenphon et al., 2012). While others have analyzed 1000s of *S. enterica* genomes, the analyses were not conducted to examine the population structure. For example, in demonstrating the software program Roary, 1000 *S. Typhi* genomes were used to test the program (Page et al., 2015). Likewise, the GenomeTrackR project utilized 32 *S. enterica* genomes to identify a *S. enterica* core, which was subsequently used as the basis for genetic distance estimates for nearly 20,000 genomes (Pettengill et al., 2016).

Previous estimates placed the core-genome size of *S. enterica* at ~2800 gene families, and the pan-genome at ~10,000 gene families (Jacobsen et al., 2011). The current study identified a strict core of 1.5 Mbp, and a conserved core of 3.2 Mbp shared among 96% of the genomes, which given an average gene size of 1000 bp is ~1500 and ~3200 genes respectively, with a much larger pan-genome of ~25,300 genes. Previous analyses found *S. enterica* to have a closed pan-genome (Jacobsen et al., 2011),

and thus the rate of discovery for new genomic regions would decrease for each new genome of the species sequenced (Tettelin et al., 2005).

In line with *S. enterica* having a closed pan-genome, when we compared it to *E. coli*, a related bacterial species with an open pan-genome (Tettelin et al., 2005), we found that the *E. coli* pan-genome was larger (37.4 Mbp), despite the fact that the *E. coli* study used less than half the number of strains in the current *Salmonella enterica* study. Additionally, more of the pan-genome of *S. enterica* was distributed among more genomes than in *E. coli* (Whiteside et al., 2016). Specifically, in *S. enterica* 70% of the pan-genome was found to belong to 100 or fewer of the genomes examined, while in *E. coli* 80% of the pan-genome was found in 100 or fewer genomes.

It should be noted that erroneously labeled, and poor quality assemblies, can greatly affect the size, analyses, and composition of the pan-genome. Software tools to evaluate assembly quality

TABLE 4 | The putative function of the *S. enterica* species-specific regions for functions that were identified more than once, utilizing the best hit for each region.

Putative protein function	Frequency
Hypothetical	64
Secreted effector	10
Membrane	7
Secretion system apparatus	5
Uncharacterized	5
Fimbrial	5
Pathogenicity island 2 effector	4
Fimbrial assembly	4
Outer membrane usher	4
mfs transporter	3
Oxidoreductase	3
Histidine kinase	3
Putative inner membrane	3
Putative cytoplasmic	3
lysr family transcriptional regulator	3
Transcriptional regulator	2
Permease	2
Outer membrane	2
Type III secretion	2
Phosphoglycerate transport	2
arac family transcriptional regulator	2
Conserved hypothetical	2
Methyl-accepting chemotaxis	2
Hybrid sensor histidine kinase/response regulator	2
Glycosyl transferase, partial	2
Phenylacetalddehyde dehydrogenase	2
Pathogenicity island 1 effector	2
n-Acetylneuramini acid mutarotase, partial	2
Type III secretion system	2
Transcriptional regulator, partial	2
Cytoplasmic	2
Fimbrial chaperone	2
Putative sialic acid transporter	2

The complete list of all putative functions is available as Supplementary File 3.

TABLE 5 | The number of subspecies-specific pan-genome markers that were universally present or absent among members of the subspecies, and not absent or present among genomes from any other subspecies.

Subspecies	No. markers
arizonae	207
diarizonae	93
enterica	9
houtenae	134
indica	192
salamae	135

have been created to help researchers identify bad data. These include QUAST (Gurevich et al., 2013), which summarizes the assembly statistics including average contig size and number of contigs; as well as CGAL (Rahman and Pachter, 2013), which uses a likelihood approach to infer assembly quality rather than

TABLE 6 | The number of pan-genome regions that were universally present and absent, as well as statistically over- or under-represented in comparison to all other genomes, within the 10 most abundant serovars within the 4870 subspecies *enterica* genomes of this study.

Serovar	No. universally present	No. universally absent
Typhi	288	2720
Typhimurium	41	698
Enteritidis	18	440
Heidelberg	121	840
Paratyphi	65	202
Kentucky	177	331
Agona	161	638
Weltevreden	426	608
Bareilly	87	436
Newport	226	360

summary statistics. As demonstrated in the current study, having a known set of species-specific genome regions can facilitate rapid quality assessment and filtering of genome assemblies. Others have proposed whole-genome MLST for this purpose as well (Babenko et al., 2016; Yoshida et al., 2016b), but the benefit of a pan-genome analysis is that it is schema free, requiring no agreed upon reference set or central repository of alleles.

S. enterica Species-Specific Regions

Previous studies have identified gene targets that are useful in the identification of *Salmonella*. These include the *fimA* gene (Cohen et al., 1996), *hilA* (Guo et al., 2000), *invA* (Malorny et al., 2003), *ttr* (Malorny et al., 2004), and *ssaN* (Chen et al., 2010). Other markers, and combinations thereof have been developed for use in RT-PCR (Postollec et al., 2011), and other detection platforms such as loop-mediated isothermal amplification (Kokkinos et al., 2014). Additionally, the identification of serovar based on allelic variation in somatic and flagellar genes has previously been conducted, with at least four laboratory methods currently available [the *Salmonella* genoserotyping assay (Yoshida et al., 2014), and the commercial assays: *Salmonella* Serogenotyping Assay, Check&Trace *Salmonella*, and xMAP *Salmonella* serotyping assay], capable of identifying over 100 of the most common *Salmonella enterica* serovars in some cases (Yoshida et al., 2016a). The recently released software, the *Salmonella* in silico typing resource (SISTR), is capable of providing *Salmonella* serovar prediction from WGSs for 90% (2,190) of all serovars (Yoshida et al., 2016b).

Despite the utility of the previously mentioned methods, previous marker-discovery studies have used at most 100s of *Salmonella* strains, while the current study examines nearly 5000. Further, the current study analyzes the entire pan-genome for predictive markers, and identified over 400 that were specific to the species, as well as others being predictive for both subspecies and serovar.

The host intestinal environment consists of a multitude of bacterial species competing for scarce nutritional sources such as carbohydrates, direct antagonistic competition with other bacterial cells, and competition for access to the host

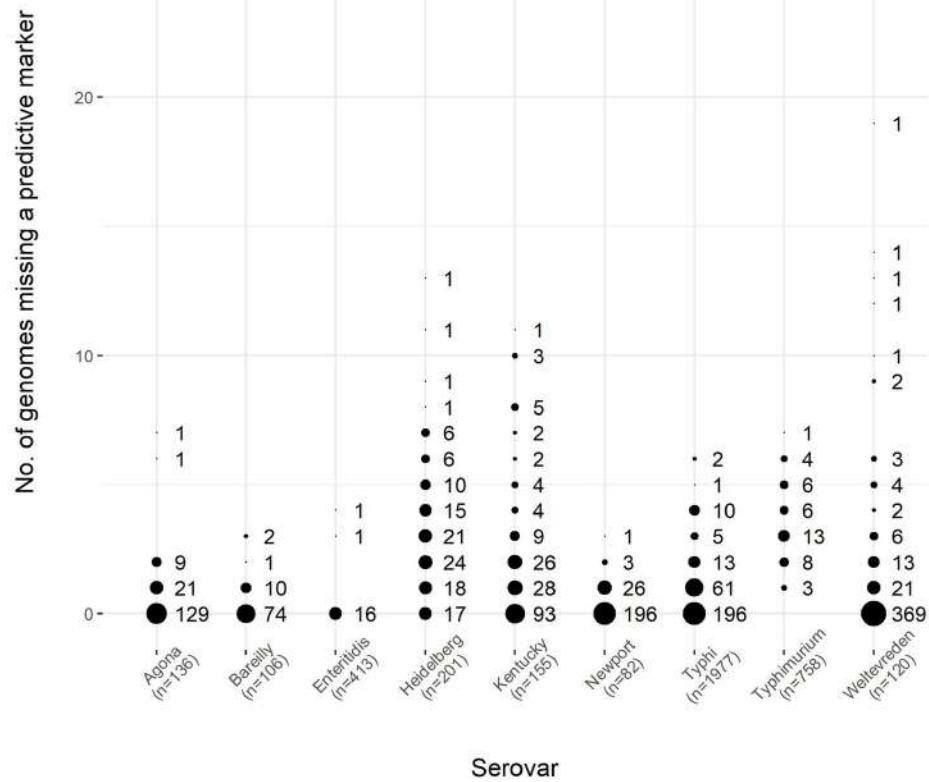


FIGURE 7 | The number of predictive markers from the GenBank dataset found within the Enterobase dataset for nine serovars of *S. enterica*, which encompassed a test set of 3948 genomes. The number of genomes for each serovar was the same between the GenBank and Enterobase datasets, as shown in **Table 6**. The size of the circles is proportional to the number of predictive markers from the GenBank dataset found in the Enterobase dataset. The number of genomes for each serovar is given in the horizontal axis label. Using serovar Agona as an example, there were 136 genomes in both the GenBank and Enterobase datasets, and 129 of the 161 predictive markers from the GenBank dataset were found in all of the genomes from the Enterobase dataset, whereas 21 of the GenBank predictive markers were found in all but one (135) of the Enterobase genomes examined.

intestine, where stable attachment and colonization of the local environment are possible (Sana et al., 2016). The normal intestinal microflora offer protection to the host against enteric pathogens such as *S. enterica*, but disruption of the intestinal environment by virulence factors and effector proteins secreted by the pathogen itself, or external factors including antibiotics, have been shown to alter the composition of the microbiota, and allow pathogens such as *S. enterica* to proliferate (Ng et al., 2013).

Nutritional competition exists for free metabolic compounds, such as carbohydrates that are readily available, as well as others that are sequestered in forms such as the intestinal mucus, which is composed of sialic sugar acids (McDonald et al., 2016). In the gut, these sugar acids exist as a conjugate in the alpha form, which is useful for bacteria such as *Salmonella*, need to be converted to the beta form by a mutarotase enzyme (Severi et al., 2008). In this study, we identified n-acetylneurameric acid mutarotase genes as species-specific genomic regions, along with

sialic acid transporter genes. It is possible the presence of these systems allow *S. enterica* to more efficiently compete with the host microbiota by efficiently utilizing scarce metabolic sources.

It was also previously found that sialic acid on the surface of host colon cells increased colonization by *S. Typhi*, and disialylation of these cells reduced the adherence of the *Salmonella* strains by 41% (Sakarya et al., 2010). This was also demonstrated in *S. Typhimurium*, where following antibiotic treatment, the presence of free sialic acid increased, and the ability to utilize it was correlated with higher levels of bacterial colonization of the host gut (Ng et al., 2013).

Enzymes that utilize sialic acids have previously been shown to be present in 452 bacterial species, including other pathogens such as *Vibrio cholerae*, but the genomic regions found in the current study were sufficiently unique at the nucleotide level to be determinative for *S. enterica* (McDonald et al., 2016).

In addition to species-specific regions used to gain a metabolic advantage, a number of secretion system and effector proteins were identified as diagnostic of *S. enterica*. These included components of the Type VI secretion system (T6SS), which is a contact-dependent, syringe-like secretion system that allows *S. enterica* to directly kill other competing bacteria that it comes into physical contact with (Brunet et al., 2015), and is encoded on the *Salmonella* Pathogenicity Island 6 (Sana et al., 2016). It has been demonstrated that silencing the T6SS via H-NS repression (histone-like nucleoid structuring), reduces inter-bacterial killing of *S. enterica* (Brunet et al., 2015). It was also previously shown that commensal bacteria are killed by *S. enterica* in a T6SS-dependent manner, that the T6SS was required for *Salmonella* to establish infection in the host gut, and that increased concentrations of bile salts resulted in a concomitant increase in T6SS anti-bacterial activity (Sana et al., 2016). The T6SS itself has been shown to have been independently acquired from four separate lineages within five of the six *S. enterica* subspecies (Desai et al., 2013).

Like the T6SS, the type III secretion system (T3SS) found within *S. enterica* is a syringe like apparatus that injects effector proteins into host cells (Kubori et al., 2000). There are two T3SS found within *S. enterica*: the first is encoded on the *Salmonella* Pathogenicity Island 1 (SPI1) and is required for invasion of host cells; the second is encoded on *Salmonella* Pathogenicity Island 2 (SPI2), and is required for survival and proliferation within the host macrophage cells (Hensel et al., 1998; Bijlsma and Groisman, 2005). The innate host immune system utilizes the inflammatory response to help reduce the proliferation of bacterial pathogens (Sun et al., 2016). *S. enterica* has developed a means of regulating host inflammation via the SPI1 T3SS, whereby secreted effector proteins target the NF-κB signaling pathway, reduce inflammation and host tissue damage, and allow increased *S. enterica* propagation within the host. *S. enterica* also relies on free long-chain fatty acids within the host to regulate T3SS expression, and provide a cue to the bacteria to up-regulate genes necessary for host intestinal colonization (Golubeva et al., 2016).

The current study identified many secretion system and effector proteins as being species-specific, as well as proteins for attachment to the host, such as fimbriae. These proteins allow *S. enterica* to compete within the intestinal environment, and take up residence within the host, where it can proliferate.

Effector proteins and other virulence factors aid in the colonization of the host, and are frequently horizontally acquired and are present on mobile elements such as integrated bacteriophages (Moreno Switt et al., 2013). Previous work identified clusters of phages that carried virulence factors such as adhesins and antimicrobial resistance determinants within *S. enterica* (Moreno Switt et al., 2013).

Additionally, many of the genes associated with bacteriophage in *S. enterica* have been found to be of the putative and hypothetical class (Penadés et al., 2015). The current study identified a large accessory gene pool that contained many hypothetical and putative genes, which were also the most abundant category of species-specific genomic regions. The proteins of putative and unknown function may aid in colonizing

warm-blooded animals, or specific animal or environmental niches. Previous studies identified genotype/phenotype correlations of *S. Typhimurium* that had particular gene complements associated with specific food sources (Hayden et al., 2016). The same study also postulated that specific phage repertoires may give phylogenetically distant strains a similar accessory gene content, and therefore similar niche specificity. Previously, 285 gene families were identified as being recruited into *S. enterica*, where most of these genes had unknown function, but were postulated to be important for its survival and infection of its host (Desai et al., 2013). It is therefore not surprising to find that the most abundant species-specific category of genomic regions are those of unknown or putative function; they likely represent genes enhancing the ability of *S. enterica* to propagate within warm-blooded animals, but they have not yet been fully characterized. The other genomic regions diagnostic of *S. enterica* include means for disseminating these fitness genes within the population, competing for resources in the host, and attaching and proliferating. The *S. enterica* species-specific regions likely give a good overview of the factors responsible for making it such an effective pathogen and intestinal inhabitant.

Specific Regions for Subspecies and Serovar

The current study recapitulates the phylogenetic relationship of the six *S. enterica* subspecies that has been previously described by others (Desai et al., 2013). However, the number of species-specific regions found within each subspecies does not follow the same pattern. For example, *diarizonae* is more distantly related to *enterica* than subspecies *indica*, but contains more species-specific regions, and the branch lengths on the tree are shorter. This indicates that although the *diarizonae* strains diverged longer ago than the *houtenae* strains, they have accumulated less genomic change. Both subspecies *diarizonae* and *houtenae* strains are associated with reptile-acquired salmonellosis (Schroter et al., 2004; Horvath et al., 2016), but the differences in genomic change may reflect the specific reptile niches that each inhabit.

Genomic regions specific to each subspecies were identified, the presence of which were unambiguously indicative of each subspecies. The most abundant subspecies in the current analyses, *enterica*, had the fewest specific markers present (9), while the most distantly related subspecies *arizonae*, had the most specific markers (207). These results indicate that just as core genome size decreases with the number of genomes examined, so too do the number of markers “core” to each subspecies. As more genomes in subspecies *arizonae* and closely related subspecies are examined, we would expect fewer genomic regions to remain specific for the subspecies. This has important implications for designing a set of markers indicative for subspecies, indicating that a group of redundant markers should be used, and that a sampling of the diversity within a subspecies is first required to identify genomic regions that are truly core.

This was also observed within serovar for subspecies *enterica* strains. The original study examining the pan-genome of *S. enterica* used a set of 45 genomes and was able to identify

unique gene families for each serovar examined, with Enteritidis having the fewest (29), and Typhi having the most (349) (Jacobsen et al., 2011). The results of the current study showed no unique genomic regions for any of the serovars with a sample set of 4893 quality filtered genomes. Although genomic regions universally present for each serovar were observed, and followed the same pattern with Enteritidis having the fewest (18), and Typhi having the most (288), these regions were also observed among genomes of other serovars, even though they were statistically over-represented for the serovar in question. The presence of these predictive markers in nearly all of the genomes within the Enterobase test dataset indicates that the markers are robust, indicative of serovar, and could be combined to determine the likelihood of a genome being of a particular serovar.

When examining the average number of the 404 species-specific regions found among the *enterica* serovars, it was interesting to observe that Enteritidis, which had the fewest number of universal genomic regions, had the highest average number of species-specific regions; likewise Typhi, which had the most universally shared genomic regions, had one of the lowest averages of species-specific regions present. These results indicate that Enteritidis is the serovar that is closest to being the “core” example of a *S. enterica* genome, while Typhi is the serovar that is the most divergent. *S. Enteritidis* is the most common cause of enteric *Salmonella* infection, causing upward of one quarter of all infections, and is prevalent in chickens as well as their eggs (Chai et al., 2012). Conversely, *S. Typhi* is a human adapted serovar, responsible for Typhoid fever, and observed to have undergone genome degradation, rearrangement, and acquisition through horizontal gene-transfer, as it has evolved within its human host (Sabbagh et al., 2010; Klemm et al., 2016). It thus appears that genomic change enabling adaptation to a host creates a genomic pool that distinguishes a group from others of the same species. At the same time, genetically similar serovars that maintain a broad host range do not undergo as much selection for genomic change are much harder to distinguish as separate groups, but much easier to identify as members of the subspecies.

Core and Pan-genome Comparison

Most phylogenetic studies focus on variation within homologs in the core genome to infer evolutionary relationships (Treangen et al., 2014), as paralogs and horizontally transferred elements confound the evolutionary signal found in genes obtained through vertical descent over time (Gabaldón and Koonin, 2013). While this approach is undoubtedly useful for long-term evolutionary analyses, when attempting to identify phenotypic linkages between phylogenetic clades, the accessory genome needs to be taken into account, as non-ubiquitous genomic regions allow different groups within the species to occupy and thrive in specific niches (Polz et al., 2013). Additionally, it has recently been shown that regulatory switching to non-homologous regulatory regions acquired via horizontal gene transfer happens in many bacteria (Oren et al., 2014). It was further shown that regulatory regions can move without the genes they regulate moving, and that at least 16% of the differences in expression observed within an *E. coli* population were explained by this regulatory switching.

It is therefore prudent to examine both the accessory genome, and not just genes, but non-coding DNA as well, as both have been shown to influence gene expression, and niche specificity. Recent studies have shown that the concordance between a phylogeny based on core genome SNPs and the presence/absence of pan-genome regions is high. For example, in a study examining *E. coli* lineage ST131, the core and accessory genomes showed high concordance, and the combined analyses of both allowed the analyses of the evolution of the *E. coli* lineage at a resolution not possible if only a restricted portion of the genome had been considered (McNally et al., 2016). The current study shows the same concordant relationship within *S. enterica* between the core and accessory genome, indicating that the accessory genome is not just randomly acquired genomic material, but that selection within specific niches establishes a complement of genes and regulatory elements that enable the survival of the *S. enterica* strains present. It also suggests that to understand why particular clades are more virulent, or possess a particular phenotype, a pan-genomic approach should be used in comparative analyses.

CONCLUSION

We examined a quality filtered set of 4893 genomes, the largest pan-genomic study of the *S. enterica* species to date. We identified a pan-genome of 25.3 Mbp, a strict core of 1.5 Mbp present in all genomes, and a conserved core of 3.2 Mbp found in at least 96% of the genomes in this study. In addition we identified 404 species-specific regions, within which a minimum set of two was required to unambiguously identify a genome as being part of the species *S. enterica*. These species-specific regions were found to have functions related to the propagation in and colonization of the host, including the utilization of sialic acid in intestinal mucus, secretion systems for attachment to the host, and the killing of other host microbiota. Within subspecies *enterica*, the species-specific regions were found most frequently in serovar Enteritidis. Each of the six subspecies was found to have genomic regions specific to it; however, the number of subspecies-specific regions appeared to be correlated with the level of sampling of the diversity within the subspecies. No serovar had pan-genome regions that were present in all of its representative genomes and absent in all other serovar genomes; however, each serovar did have genomic regions that were universally present among all constituent members, and statistically predictive of the serovar. *S. Typhi*, which is host-adapted to humans, was found to have the most universal markers predictive of its serovar. The phylogeny based on SNPs within the conserved core genome was found to be highly concordant to that produced by a phylogeny using the presence/absence of the entire pan-genome, and both agreed with phylogenies previously reported for *S. enterica*. Together, the core and accessory genome offered a more complete picture of the diversity within the genomes than either alone. The genomic regions identified in this study that are predictive of the species *S. enterica*, its six subspecies, and the serovar groups within subspecies *enterica*, could be developed into simple and rapid diagnostic tests, with uses ranging from food safety to public health. Additionally, the tools

and methods described in this study could be generally applicable as a pan-genomics framework for future population studies, or those looking for genotype/phenotype linkages.

AUTHOR CONTRIBUTIONS

CL: designed the experiments, analyzed the data, and wrote the manuscript. MW: designed the experiments, and wrote the manuscript. VG: designed the experiments, and wrote the manuscript.

REFERENCES

- Ananensen, D. M., Feil, E. J., Holden, M. T. G., Dordel, J., Yeats, C. A., Fedosejev, A., et al. (2016). Whole-genome sequencing for routine pathogen surveillance in public health: population snapshot of invasive *Staphylococcus aureus* in Europe. *mBio* 7:e00444-16. doi: 10.1128/mBio.00444-16
- Allard, M. W., Luo, Y., Strain, E., Li, C., Keys, C. E., Son, I., et al. (2012). High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. *BMC Genomics* 13:32. doi: 10.1186/1471-2164-13-32
- Ashton, P. M., Nair, S., Peters, T. M., Bale, J. A., Powell, D. G., Painset, A., et al. (2016). Identification of *Salmonella* for public health surveillance using whole genome sequencing. *PeerJ* 4:e1752. doi: 10.7717/peerj.1752
- Babenko, D., Azizov, I., and Toleman, M. (2016). wgMLST as a standardized tool for assessing the quality of genome assembly data. *Int. J. Infect. Dis.* 45:329. doi: 10.1016/j.ijid.2016.02.714
- Bergholz, T. M., Moreno Switt, A. I., and Wiedmann, M. (2014). Omics approaches in food safety: Fulfilling the promise? *Trends Microbiol.* 22, 275–281. doi: 10.1016/j.tim.2014.01.006
- Bijlsma, J. J., and Groisman, E. A. (2005). The PhoP/PhoQ system controls the intramacrophage type three secretion system of *Salmonella enterica*. *Mol. Microbiol.* 57, 85–96. doi: 10.1111/j.1365-2958.2005.04668.x
- Brunet, Y. R., Khodr, A., Logger, L., Aussel, L., Mignot, T., Rimsky, S., et al. (2015). H-NS silencing of the *Salmonella* pathogenicity island 6-encoded type VI secretion system limits *Salmonella enterica* serovar typhimurium interbacterial killing. *Infect. Immun.* 83, 2738–2750. doi: 10.1128/IAI.00198-15
- Chai, S. J., White, P. L., Lathrop, S. L., Solghan, S. M., Medus, C., McGlinchey, B. M., et al. (2012). *Salmonella enterica* serotype enteritidis: increasing incidence of domestically acquired infections. *Clin. Infect. Dis.* 54(Suppl. 5), S488–S497. doi: 10.1093/cid/cis231
- Chen, J., Zhang, L., Paoli, G. C., Shi, C., Tu, S. I., and Shi, X. (2010). A real-time PCR method for the detection of *Salmonella enterica* from food using a target sequence identified by comparative genomic analysis. *Int. J. Food Microbiol.* 137, 168–174. doi: 10.1016/j.ijfoodmicro.2009.12.004
- Cohen, H., Mechanda, S., and Lin, W. (1996). PCR amplification of the fimA gene sequence of *Salmonella* typhimurium, a specific method for detection of *Salmonella* spp. *Appl. Environ. Microbiol.* 62, 4303–4308.
- den Bakker, H. C., Allard, M. W., Bopp, D., Brown, E. W., Fontana, J., Iqbal, Z., et al. (2014). Rapid whole-genome sequencing for surveillance of *Salmonella enterica* serovar enteritidis. *Emerg. Infect. Dis.* 20, 1306–1314. doi: 10.3201/eid2008.131399
- Deng, X., den Bakker, H. C., and Hendriksen, R. S. (2016). Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu. Rev. Food Sci. Technol.* 7, 353–374. doi: 10.1146/annurev-food-041715-033259
- Desai, P. T., Porwollik, S., Long, F., Cheng, P., Wollam, A., Bhonagiri-Palsikar, V., et al. (2013). Evolutionary genomics of *Salmonella enterica* subspecies. *mBio* 4:e00579-12. doi: 10.1128/mBio.00579-12
- Franz, E., Gras, L. M., and Dallman, T. (2016). Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. *Curr. Opin. Food Sci.* 8, 74–79. doi: 10.1016/j.cofs.2016.04.004
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565
- Gabaldón, T., and Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* 14, 360–366. doi: 10.1038/nrg3456
- Gal-Mor, O., Boyle, E. C., and Grassl, G. A. (2014). Same species, different diseases: how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. *Front. Microbiol.* 5:391. doi: 10.3389/fmicb.2014.00391
- Golubeva, Y. A., Ellermeier, J. R., Chubiz, J. E. C., and Slauach, J. M. (2016). Intestinal long-chain fatty acids act as a direct signal to modulate expression of the *Salmonella* pathogenicity island 1 type III secretion system. *mBio* 7:e02170-15. doi: 10.1128/mBio.02170-15
- Guo, X., Chen, J., Beuchat, L. R., and Robert, E. (2000). PCR detection of *Salmonella enterica* serotype montevideo in and on raw tomatoes using primers derived from hilA. *Appl. Environ. Microbiol.* 66, 5248–5252. doi: 10.1128/AEM.66.12.5248-5252.2000. Updated
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086
- Hayden, H. S., Matamouros, S., Hager, K. R., Brittnacher, M. J., Rohmer, L., Raday, M. C., et al. (2016). Genomic analysis of *Salmonella enterica* serovar Typhimurium characterizes strain diversity for recent U.S. salmonellosis cases and identifies mutations linked to loss of fitness under nitrosative and oxidative stress. *mBio* 7:e00154-16. doi: 10.1128/mBio.00154-16
- Hensel, M., Shea, J. E., Waterman, S. R., Mundy, R., Nikolaus, T., Banks, G., et al. (1998). Genes encoding putative effector proteins of the type III secretion system of *Salmonella* pathogenicity island 2 are required for bacterial virulence and proliferation in macrophages. *Mol. Microbiol.* 30, 163–174. doi: 10.1046/j.1365-2958.1998.01047.x
- Horvath, L., Kraft, M., Fostopoulos, K., Falkowski, A., and Tarr, P. E. (2016). *Salmonella enterica* subspecies *difarionae* maxillary sinusitis in a snake handler: first report. *Open Forum Infect. Dis.* 3:ofw066. doi: 10.1093/ofid/ofw066
- Jacobsen, A., Hendriksen, R. S., Aarestrup, F. M., Ussery, D. W., and Friis, C. (2011). The *Salmonella enterica* Pan-genome. *Microb. Ecol.* 62, 487–504. doi: 10.1007/s00248-011-9880-1
- Kirk, M. D., Pires, S. M., Black, R. E., Caipo, M., Crump, J. A., Devleesschauwer, B., et al. (2015). World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. *PLoS Med.* 12:e1001921. doi: 10.1371/journal.pmed.1001921
- Klemm, E. J., Gkrania-Klotsas, E., Hadfield, J., Forbester, J. L., Harris, S. R., Hale, C., et al. (2016). Emergence of host-adapted *Salmonella* Enteritidis through rapid evolution in an immunocompromised host. *Nat. Microbiol.* 1:15023. doi: 10.1038/NMICROBIOL.2015.23
- Kokkinos, P. A., Ziros, P. G., Bellou, M., and Vantarakis, A. (2014). Loop-Mediated Isothermal Amplification (LAMP) for the detection of *Salmonella* in food. *Food Anal. Methods* 7, 512–526. doi: 10.1007/s12161-013-9748-8
- Kubori, T., Sukhan, A., Aizawa, S. I., and Galán, J. E. (2000). Molecular characterization and assembly of the needle complex of the *Salmonella*

ACKNOWLEDGMENT

Thanks to Peter Kruczakiewicz of the Public Health Agency of Canada for providing the curated metadata for the Enterobase genomes used in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.01345/full#supplementary-material>

- typhimurium type III protein secretion system. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10225–10230. doi: 10.1073/pnas.170128997
- Laing, C., Buchanan, C., Taboada, E. N., Zhang, Y., Kropinski, A., Villegas, A., et al. (2010). Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics* 11:461. doi: 10.1186/1471-2105-11-461
- Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aarestrup, F. M., and Ussery, D. W. (2012). Genomic variation in *Salmonella enterica* core genes for epidemiological typing. *BMC Genomics* 13:88. doi: 10.1186/1471-2164-13-88
- Levine, M. M., Stinear, T., Holt, K. E., Robins-Browne, R. M., Ingle, D. J., Kuzevski, A., et al. (2016). *In silico* serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb. Genomics* 2:e000064. doi: 10.1099/mgen.0.000064
- Lupolova, N., Dallman, T. J., Matthews, L., Bono, J. L., and Gally, D. L. (2016). Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proc. Natl. Acad. Sci. U.S.A.* 113, 11312–11317. doi: 10.1073/pnas.1606567113
- Majowicz, S. E., Musto, J., Scallan, E., Angulo, F. J., Kirk, M., O'Brien, S. J., et al. (2010). The global burden of nontyphoidal *Salmonella* gastroenteritis. *Clin. Infect. Dis.* 50, 882–889. doi: 10.1086/650733
- Malorny, B., Hoorfar, J., Bunge, C., and Helmuth, R. (2003). Multicenter validation of the analytical accuracy of *Salmonella* PCR: towards an international standard. *Appl. Environ. Microbiol.* 69, 290–296. doi: 10.1128/AEM.69.1.290
- Malorny, B., Paccassoni, E., Fach, P., Martin, A., Helmuth, R., and Bunge, C. (2004). Diagnostic real-time PCR for detection of *Salmonella* in food. *Appl. Environ. Microbiol.* 70, 7046–7052. doi: 10.1128/AEM.70.12.7046
- McDermott, P. F., Tyson, G. H., Kabera, C., Chen, Y., Li, C., Folster, J. P., et al. (2016). The use of whole genome sequencing for detecting antimicrobial resistance in nontyphoidal *Salmonella*. *Antimicrob. Agents Chemother.* 60, 5515–5520. doi: 10.1128/AAC.01030-16
- McDonald, N. D., Lubin, J. B., Chowdhury, N., and Boyd, E. F. (2016). Host-derived sialic acids are an important nutrient source required for optimal bacterial fitness in vivo. *mBio* 7:e02237-15. doi: 10.1128/mBio.02237-15
- McNally, A., Oren, Y., Kelly, D., Pascoe, B., Dunn, S., Sreecharan, T., et al. (2016). Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet.* 12:e1006280. doi: 10.1371/journal.pgen.1006280
- Moreno Switt, A. I., Orsi, R. H., den Bakker, H. C., Vongkamjan, K., Altier, C., and Wiedmann, M. (2013). Genomic characterization provides new insight into *Salmonella* phage diversity. *BMC Genomics* 14:481. doi: 10.1186/1471-2164-14-481
- Ng, K. M., Ferreyra, J. A., Higginbottom, S. K., Lynch, J. B., Kashyap, P. C., Gopinath, S., et al. (2013). Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. *Nature* 502, 96–99. doi: 10.1038/nature12503
- Oprijnen, T. V., Camilli, A., Oprijnen, T. V., and Camilli, A. (2012). A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome Res.* 22, 2541–2551. doi: 10.1101/gr.137430.112
- Oren, Y., Smith, M. B., Johns, N. I., Kaplan Zeevi, M., Biran, D., Ron, E. Z., et al. (2014). Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 111, 16112–16117. doi: 10.1073/pnas.1413272.2011
- Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421
- Penadés, J. R., Chen, J., Quiles-Puchalt, N., Carpina, N., and Novick, R. P. (2015). Bacteriophage-mediated spread of bacterial virulence genes. *Curr. Opin. Microbiol.* 23, 171–178. doi: 10.1016/j.mib.2014.11.019
- Pettengill, J. B., Pightling, A. W., Baugher, J. D., Rand, H., and Strain, E. (2016). Real-time pathogen detection in the era of whole-genome sequencing and big data: comparison of k-mer and site-based methods for inferring the genetic distances among tens of thousands of *Salmonella* samples. *PLoS ONE* 11:e0166162. doi: 10.1371/journal.pone.0166162
- Polz, M. F., Alm, E. J., and Hanage, W. P. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* 29, 170–175. doi: 10.1016/j.tig.2012.12.006
- Postollec, F., Falentin, H., Pavan, S., Combrisson, J., and Sohier, D. (2011). Recent advances in quantitative PCR (qPCR) applications in food microbiology. *Food Microbiol.* 28, 848–861. doi: 10.1016/j.fm.2011.02.008
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rahman, A., and Pachter, L. (2013). CGAL: computing genome assembly likelihoods. *Genome Biol.* 14:R8. doi: 10.1186/gb-2013-14-1-r8
- Sabbagh, S. C., Forest, C. G., Lepage, C., Leclerc, J. M., and Daigle, F. (2010). So similar, yet so different: uncovering distinctive features in the genomes of *Salmonella enterica* serovars Typhimurium and Typhi. *FEMS Microbiol. Lett.* 305, 1–13. doi: 10.1111/j.1574-6968.2010.01904.x
- Sakarya, S., Göktürk, C., Özürk, T., and Ertugrul, M. B. (2010). Sialic acid is required for nonspecific adherence of *Salmonella enterica* ssp. *enterica* serovar Typhi on Caco-2 cells. *FEMS Immunol. Med. Microbiol.* 58, 330–335. doi: 10.1111/j.1574-695X.2010.00650.x
- Sana, T. G., Flaugnatti, N., Lugo, K. A., Lam, L. H., Jacobson, A., Baylot, V., et al. (2016). *Salmonella* Typhimurium utilizes a T6SS-mediated antibacterial weapon to establish in the host gut. *Proc. Natl. Acad. Sci. U.S.A.* 113, E5044–E5051. doi: 10.1073/pnas.1608858113
- Scallan, E., Hoekstra, R. M., Mahon, B. E., Jones, T. F., and Griffin, P. M. (2015). An assessment of the human health impact of seven leading foodborne pathogens in the United States using disability adjusted life years. *Epidemiol. Infect.* 143, 2795–2804. doi: 10.1017/S0950268814003185
- Scharff, R. L., Besser, J., Sharp, D. J., Jones, T. F., Peter, G. S., and Hedberg, C. W. (2016). An economic evaluation of PulseNet: a network for foodborne disease surveillance. *Am. J. Prev. Med.* 50, S66–S73. doi: 10.1016/j.amepre.2015.09.018
- Schroter, M., Roggentin, P., Hofmann, J., Speicher, A., Laufs, R., and Mack, D. (2004). Pet snakes as a reservoir for *Salmonella enterica* subsp. *diarizonae* (serogroup IIIb): a prospective study. *Appl. Environ. Microbiol.* 70, 613–615. doi: 10.1128/AEM.70.1.613–615.2004
- Severi, E., Müller, A., Potts, J. R., Leech, A., Williamson, D., Wilson, K. S., et al. (2008). Sialic acid mutarotation is catalyzed by the *Escherichia coli* β -propeller protein YhhT. *J. Biol. Chem.* 283, 4841–4849. doi: 10.1074/jbc.M707822200
- Sheppard, S. K., Jolley, K. A., and Maiden, M. C. J. (2012). A gene-by-gene approach to bacterial population genomics: whole genome MLST of *Campylobacter*. *Genes* 3, 261–277. doi: 10.3390/genes3020261
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033
- Sun, H., Kamanova, J., Lara-Tejero, M., and Galán, J. E. (2016). A Family of *Salmonella* type III secretion effector proteins selectively targets the NF- κ B signaling pathway to preserve host homeostasis. *PLoS Pathog.* 12:e1005484. doi: 10.1371/journal.ppat.1005484
- Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102
- Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15:524. doi: 10.1186/s13059-014-0524-x
- Tyson, G. H., McDermott, P. F., Li, C., Chen, Y., Tadesse, D. A., Mukherjee, S., et al. (2015). WGS accurately predicts antimicrobial resistance in *Escherichia coli*. *J. Antimicrob. Chemother.* 70, 2763–2769. doi: 10.1093/jac/dkv186
- Waryah, C. B., Gogoï-Tiwari, J., Wells, K., Eto, K. Y., Masoumi, E., Costantino, P., et al. (2016). Diversity of virulence factors associated with West Australian methicillin-sensitive *Staphylococcus aureus* isolates of human origin. *BioMed Res. Int.* 2016:8651918. doi: 10.1155/2016/8651918
- Whiteside, M. D., Laing, C. R., Manji, A., Krucziewicz, P., Taboada, E. N., and Gannon, V. P. J. (2016). SuperPhy: predictive genomics for the bacterial pathogen *Escherichia coli*. *BMC Microbiol.* 16:65. doi: 10.1186/s12866-016-0680-0
- Yoshida, C., Gurnik, S., Ahmad, A., Blimkie, T., Murphy, S. A., Kropinski, A. M., et al. (2016a). Evaluation of molecular methods for identification of *Salmonella* serovars. *J. Clin. Microbiol.* 54, 1992–1998. doi: 10.1128/JCM.00262-16
- Yoshida, C., Krucziewicz, P., Laing, C. R., Lingohr, E. J., Gannon, V. P. J., Nash, J. H. E., et al. (2016b). The *Salmonella* *In silico* typing resource (SISTR): an

- open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS ONE* 11:e0147101. doi: 10.1371/journal.pone.0147101
- Yoshida, C., Lingohr, E. J., Trognitz, F., MacLaren, N., Rosano, A., Murphy, S. A., et al. (2014). Multi-laboratory evaluation of the rapid genosotyping array (SGSA) for the identification of *Salmonella* serovars. *Diagn. Microbiol. Infect. Dis.* 80, 185–190. doi: 10.1016/j.diagmicrobio.2014.08.006
- Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2016). GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628
- Zhao, S., Tyson, G. H., Chen, Y., Li, C., Mukherjee, S., Young, S., et al. (2016). Whole-genome sequencing analysis accurately predicts antimicrobial resistance phenotypes in *Campylobacter* spp. *Appl. Environ. Microbiol.* 82, 459–466. doi: 10.1128/AEM.02873-15

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Laing, Whiteside and Gannon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Prophage Integrase Typing Is a Useful Indicator of Genomic Diversity in *Salmonella enterica*

Anna Colavecchio^{1*}, Yasmin D'Souza¹, Elizabeth Tompkins¹, Julie Jeukens², Luca Freschi², Jean-Guillaume Emond-Rheault², Irena Kukavica-Ibrulj², Brian Boyle², Sadjia Bekal³, Sandeep Tamber⁴, Roger C. Levesque² and Lawrence D. Goodridge^{1*}

¹ Food Safety and Quality Program, Department of Food Science and Agricultural Chemistry, McGill University, Sainte-Anne-de-Bellevue, QC, Canada, ² Institut de Biologie Intégrative et des Systèmes, Université Laval, Quebec City, QC, Canada, ³ Pathogènes entériques et Bioterrorisme, Laboratoire de santé publique du Québec, Sainte-Anne-de-Bellevue, QC, Canada, ⁴ Salmonella Research Laboratory, Bureau of Microbial Hazards, Health Canada, Ottawa, ON, Canada

OPEN ACCESS

Edited by:

Sabah Bidawid,
Health Canada, Canada

Reviewed by:

Beatrix Stessl,

Veterinärmedizinische Universität
Wien, Austria
Mehrdad Mark Tajkarimi,
Joint School of Nanoscience
and Nanoengineering, University
of North Carolina at Greensboro,
United States

*Correspondence:

Lawrence D. Goodridge
lawrence.goodridge@mcgill.ca
Anna Colavecchio
anna.colavecchio@mail.mcgill.ca

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 01 April 2017

Accepted: 26 June 2017

Published: 10 July 2017

Citation:

Colavecchio A, D'Souza Y, Tompkins E, Jeukens J, Freschi L, Emond-Rheault J-G, Kukavica-Ibrulj I, Boyle B, Bekal S, Tamber S, Levesque RC and Goodridge LD (2017) Prophage Integrase Typing Is a Useful Indicator of Genomic Diversity in *Salmonella enterica*. *Front. Microbiol.* 8:1283. doi: 10.3389/fmicb.2017.01283

Salmonella enterica is a bacterial species that is a major cause of illness in humans and food-producing animals. *S. enterica* exhibits considerable inter-serovar diversity, as evidenced by the large number of host adapted serovars that have been identified. The development of methods to assess genome diversity in *S. enterica* will help to further define the limits of diversity in this foodborne pathogen. Thus, we evaluated a PCR assay, which targets prophage integrase genes, as a rapid method to investigate *S. enterica* genome diversity. To evaluate the PCR prophage integrase assay, 49 isolates of *S. enterica* were selected, including 19 clinical isolates from clonal serovars (*Enteritidis* and *Heidelberg*) that commonly cause human illness, and 30 isolates from food-associated *Salmonella* serovars that rarely cause human illness. The number of integrase genes identified by the PCR assay was compared to the number of integrase genes within intact prophages identified by whole genome sequencing and phage finding program PHASTER. The PCR assay identified a total of 147 prophage integrase genes within the 49 *S. enterica* genomes (79 integrase genes in the food-associated *Salmonella* isolates, 50 integrase genes in *S. Enteritidis*, and 18 integrase genes in *S. Heidelberg*). In comparison, whole genome sequencing and PHASTER identified a total of 75 prophage integrase genes within 102 intact prophages in the 49 *S. enterica* genomes (44 integrase genes in the food-associated *Salmonella* isolates, 21 integrase genes in *S. Enteritidis*, and 9 integrase genes in *S. Heidelberg*). Collectively, both the PCR assay and PHASTER identified the presence of a large diversity of prophage integrase genes in the food-associated isolates compared to the clinical isolates, thus indicating a high degree of diversity in the food-associated isolates, and confirming the clonal nature of *S. Enteritidis* and *S. Heidelberg*. Moreover, PHASTER revealed a diversity of 29 different types of prophages and 23 different integrase genes within the food-associated isolates, but only identified four different phages and integrase genes within clonal isolates of *S. Enteritidis* and *S. Heidelberg*. These results demonstrate the potential usefulness of PCR based detection of prophage integrase genes as a rapid indicator of genome diversity in *S. enterica*.

Keywords: *Salmonella enterica*, foodborne pathogen, genome diversity, prophage integrase gene analysis, signature genes

INTRODUCTION

Salmonella enterica is a Gram-negative pathogen that infects humans and animals. *S. enterica* is divided into six sub-species on the basis of genetic content (Wain and O'Grady, 2017), and contains more than 2,500 serovars. *S. enterica* subspecies *enterica* is a major cause of enteric disease in humans and animals. The majority of illnesses caused by *S. enterica* are foodborne. Globally, *S. enterica* causes 93 million gastroenteritis cases and 150,000 deaths annually (Majowicz et al., 2010). In Canada, salmonellosis accounts for 87,510 human cases, and 17 deaths each year (Thomas et al., 2013, 2015). However, while there are more than 2,500 serovars of *S. enterica* subspecies I, the majority (75%) of all salmonellosis cases in Canada are caused by only 10 serovars. And three serovars, *S. Enteritidis* (30%), *S. Heidelberg* (15%), and *S. Typhimurium* (12%), account for 57% of all salmonellosis cases in Canada (Public Health Agency of Canada [PHAC], 2012). In the United States, where *Salmonella* accounts for 1.2 million illnesses and 450 deaths annually (Scallan et al., 2011), the top 10 serovars cause approximately 57% of illnesses. Characterizing *Salmonella* serovars into monophyletic and polyphyletic lineages is essential for linking outbreaks (Timme et al., 2013). While the implementation of Hazard Analysis and Critical Control Point (HACCP) programs in the food industry has reduced contamination of foods of animal origin, there has been increased recognition of *Salmonella* contamination associated with fresh produce, which accounts for approximately half of all fresh produce outbreaks due to bacteria in the United States and European Union (Callejón et al., 2015).

These statistics have led to much work aimed at identifying virulence and fitness markers in *S. enterica*, as well as questions regarding genome diversity. *S. enterica* genomes are highly diversified due to insertions and deletions (indels) (Zhou et al., 2013). Survival in different habitats, as evidenced by the large number of host species colonized by *S. enterica*, and the ability to successfully transmit through food and water, or directly from host to host has driven this diversity. High levels of intra-serovar diversity has also been recognized, as demonstrated by the acquisition of indels (51 prophages, 10 plasmids, and 6 integrative conjugational elements) by *S. enterica* Agona (Zhou et al., 2013).

Prophages are bacteriophages which have integrated into bacterial chromosomes, by means of an integrase gene, and they have been found to contribute to interstrain genetic variability (Brüssow et al., 2004). Bacteriophages are the most abundant organisms on earth, and it is estimated that there are 10^{31} phage particles in the biosphere (McNair et al., 2012). Phages are ubiquitous and can be found in any environment where their bacterial hosts are present. It has been estimated that there are 100 million phage species (Rohwer, 2003). As such, phages likely play a major role in defining the dynamics of microbial community structure and function. In fact, much of the diversity observed in closely related bacterial strains is a result of the incorporation of diverse prophages into the core bacterial genome (Brüssow et al., 2004). Prophages enhance bacterial

fitness by encoding many proteins important in virulence and antibiotic resistance.

Many studies have demonstrated the presence of numerous prophages within *S. enterica* (Figueroa-Bossi and Bossi, 1999; Kropinski et al., 2007). In one such study, Thomson et al. (2004) characterized prophages within the *S. enterica* serovar Typhi CT18 chromosome. *In silico* analyses were used to compare prophage regions in *S. Typhi* CT18, to prophages within 40 other *Salmonella* isolates using DNA microarray technology. The results indicated that the *S. Typhi* CT18 prophages had similarity to the lambda, Mu, P2 and P4 phage families. Other *S. Typhi* isolates also had similar prophages, supporting a clonal origin of this serovar. In contrast, distinct prophage variation was detected within a broad range of *Salmonella* serovars, suggesting that these phages may confer a level of specialization on their host. The authors concluded that prophages therefore play a crucial role in the generation of genetic diversity within *S. enterica*. This statement, and the lack of an universal phylogenetic marker for phages provides the rationale for this study in which we evaluated the use of prophage integrase genes as reliable indicators of genome diversity in *S. enterica*.

We compared two serovars (*Enteritidis* and *Heidelberg*) whose genomes are reported to be clonal, to a set of *S. enterica* isolates from diverse food sources (Table 1). These food associated isolates are considered to be "rare" because they belong to serovars that are not within the top 70 serovars that cause illness in Canada. We hypothesized that the integrase gene is associated with prophage diversity within *Salmonella*. Hence, screening integrase genes (via a PCR assay) would detect few and similar patterns of integrase genes in the non-diverse and clonal isolates of *S. Enteritidis* and *S. Heidelberg* but a larger number and diversity of integrase genes in the food associated isolates that are rare and diverse.

MATERIALS AND METHODS

Bacterial Isolates and Growth Conditions

Forty-nine *S. enterica* isolates (Table 1) were used in this study. Metadata for all *S. enterica* isolates can be found in the *Salmonella* Foodborne Syst-OMICS Database (SalFoS), which can be accessed at <https://salfos.ibis.ulaval.ca/>. In order to determine whether prophage integrase genes could be used to assess genome diversity in *S. enterica*, we chose isolates from serovars that are consistently implicated in outbreaks of salmonellosis and that are also clonal in nature (clinical isolates from serovars *Enteritidis* and *Heidelberg*) as well as isolates from rare serovars that were isolated from diverse food sources (food associated isolates). All isolates were maintained at -80°C in glycerol, and were revived by streaking the frozen culture on tryptic soy agar (TSA) (BD Biosciences, Mississauga, ON, Canada) followed by incubation at 37°C for 16 h. Isolated colonies were then inoculated in tryptic soy broth (TSB) (BD Biosciences, Mississauga, ON, Canada) and grown at 37°C for 16 h in an orbital shaker at a speed of 225 rpm.

TABLE 1 | List of *Salmonella enterica* isolates used in this study.

Taxon	Serovar	SalFoS ID	Origin	Source isolation
<i>Salmonella enterica</i>	Amager	S25	HC ^a	Animal Feed
<i>Salmonella enterica</i>	Ball	S26	HC	Shellfish – Shrimp
<i>Salmonella enterica</i>	Banana	S27	HC	Animal Feed
<i>Salmonella enterica</i>	Bergen	S28	HC	Shellfish – Shrimp
<i>Salmonella enterica</i>	Broughton	S29	HC	Poultry
<i>Salmonella enterica</i>	Canada	S30	HC	Chocolate
<i>Salmonella enterica</i>	Casablanca	S31	HC	Fish/Shellfish
<i>Salmonella enterica</i>	Chingola	S32	HC	Fish – Seaweed
<i>Salmonella enterica</i>	Cremieu	S33	HC	Fish – Frozen eel
<i>Salmonella enterica</i>	Daytona	S34	HC	Shellfish – Clams
<i>Salmonella enterica</i>	Duesseldorf	S35	HC	Poultry
<i>Salmonella enterica</i>	Elisabethville	S36	HC	Reptiles – Agamid
<i>Salmonella enterica</i>	Falkensee	S37	HC	Spices – Rice seasoning
<i>Salmonella enterica</i>	Fresno	S38	HC	Animal feed
<i>Salmonella enterica</i>	Godesberg	S39	HC	Sesame – Halawa
<i>Salmonella enterica</i>	Hull	S40	HC	Fish – Dried conch
<i>Salmonella enterica</i>	Indikan	S41	HC	Sesame – Tahini
<i>Salmonella enterica</i>	Kouka	S42	HC	Shellfish – Oysters
<i>Salmonella enterica</i>	Luciana	S43	HC	Fruit – Cantaloupe
<i>Salmonella enterica</i>	Luckenwalde	S44	HC	Cocoa beans
<i>Salmonella enterica</i>	Orientalis	S45	HC	Alfalfa seeds
<i>Salmonella enterica</i>	Pasing	S46	HC	Chocolate
<i>Salmonella enterica</i>	Solt	S47	HC	Cocoa beans
<i>Salmonella enterica</i>	Tado	S48	HC	Animal feed
<i>Salmonella enterica</i>	Taiping	S49	HC	Fish
<i>Salmonella enterica</i>	Taksony	S50	HC	Ox
<i>Salmonella enterica</i>	Tyresoe	S51	HC	Shellfish – Shrimp
<i>Salmonella enterica</i>	Wentworth	S52	HC	Fish – Cuttlefish
<i>Salmonella enterica</i>	Westhampton	S53	HC	Animal feed supplement
<i>Salmonella enterica</i>	Weston	S54	HC	Shellfish – Shrimp
<i>Salmonella enterica</i>	Enteritidis	S3	LSPQ ^b	Clinical
<i>Salmonella enterica</i>	Enteritidis	S4	LSPQ	Clinical
<i>Salmonella enterica</i>	Enteritidis	S5	LSPQ	Clinical
<i>Salmonella enterica</i>	Enteritidis	S6	LSPQ	Clinical
<i>Salmonella enterica</i>	Enteritidis	S7	LSPQ	Clinical
<i>Salmonella enterica</i>	Enteritidis	S8	LSPQ	Clinical
<i>Salmonella enterica</i>	Enteritidis	S9	LSPQ	Clinical
<i>Salmonella enterica</i>	Enteritidis	S10	LSPQ	Clinical
<i>Salmonella enterica</i>	Enteritidis	S11	LSPQ	Clinical
<i>Salmonella enterica</i>	Enteritidis	S12	LSPQ	Clinical
<i>Salmonella enterica</i>	Heidelberg	S430	LSPQ	Clinical
<i>Salmonella enterica</i>	Heidelberg	S429	LSPQ	Clinical
<i>Salmonella enterica</i>	Heidelberg	S371	LSPQ	Clinical
<i>Salmonella enterica</i>	Heidelberg	S426	LSPQ	Clinical
<i>Salmonella enterica</i>	Heidelberg	S427	LSPQ	Clinical
<i>Salmonella enterica</i>	Heidelberg	S370	LSPQ	Clinical
<i>Salmonella enterica</i>	Heidelberg	S431	LSPQ	Clinical
<i>Salmonella enterica</i>	Heidelberg	S432	LSPQ	Clinical
<i>Salmonella enterica</i>	Heidelberg	S433	LSPQ	Clinical

Further details can be found in the SalFoS database at <https://salfos.ibis.ulaval.ca/>. ^aHealth Canada, Ottawa, ON, Canada. ^bLaboratoire de santé publique du Québec, Sainte-Anne-de-Bellevue, QC, Canada.

DNA Extraction and Amplification of Bacteriophage Specific Integrase Genes

Bacterial DNA was extracted from an overnight TSB culture using the DNeasy Blood and Tissue kit (Qiagen Inc., Germantown, MD, United States) according to the manufacturer's instructions. The polymerase chain reaction (PCR) was used to amplify prophage tyrosine integrase genes using a prophage integrase assay previously described by Balding et al. (2005). Briefly, a set of 11 degenerate primer sets were designed by aligning the conserved regions, designated as "box I" and "box II," of the tyrosine integrase of 32 enteric prophages encoded by members of the *Enterobacteriaceae* family. The two conserved regions are located in the C-terminal of the tyrosine integrase and consist of residues A202-G227 ("Box I") and T206-D344 ("Box II") in the Lambda prophage. Prophages encoding an integrase gene with similar "box I" and "box II" regions were grouped into eight primer sets. Primer set 5 was further subdivided into groups 5A, 5B, and 5C, and group 6 was further subdivided into groups 6A and 6B, for a total of 11 degenerate primer sets. The primer sets and sequences can be found in Balding et al. (2005).

Polymerase chain reaction amplification was conducted in a Peltier Thermal Cycler (PTC-100, Bio-Rad Laboratories, Inc., Mississauga, ON, Canada), and commenced with DNA denaturation for 5 min at 94°C followed by 25 cycles consisting of 94°C for 30 s, 40°C for 30 s, 72°C for 30 s and a final extension at 72°C, for 7 min. PCR amplicons ranged in size from 280 to 447 bp, and were resolved by electrophoresis in 1X Tris/Borate/EDTA (TBE) buffer on 1% (w/v) agarose gels that contained 1x SYBR Safe stain (Thermo Fisher Scientific, Waltham, MA, United States). Following gel electrophoresis, amplicons were visualized under UV illumination.

Whole Genome Sequencing and Bioinformatic Analysis

Whole genome sequencing was performed at the EcoGenomics Analysis Platform (IBIS, Université Laval, Quebec City, QC, Canada) on an Illumina MiSeq using 300-bp paired-end libraries with 40× coverage. The raw reads were assembled using the A5 pipeline (Tritt et al., 2012). Each of the 49 assembled genomes were analyzed by PHASTER to identify the presence of prophages and their integrase genes (Arndt et al., 2016). Only prophages identified as "complete" or "intact" were considered for further analysis. The identity of all intact prophage sequences detected by PHASTER was confirmed by BLAST (Altschul et al., 1990).

Phylogenetic Tree Construction

Parsnp, included in the Harvest suite of core-genome alignment tools, was performed to produce a rapid core-genome alignment based on SNPs (1000 bootstraps) of the core genome sequences of the 30 food associated *Salmonella* isolates and 19 clinical *S. Enteritidis* and *S. Heidelberg* isolates (Treangen et al., 2014). The alignment data was converted to Newick format and a unrooted maximum-likelihood tree was constructed and edited with Interactive tree of life (iTOL) version 3 (Letunic and Bork, 2016). Whole genome alignments of the prophage sequences as well as their integrase genes, and construction of unrooted

maximum-likelihood trees were performed using BioNumerics version 7.6.2 (Applied Maths, 2017). All phylogenetic trees constructed using BioNumerics were converted to Newick files and edited with iTOL (Letunic and Bork, 2016).

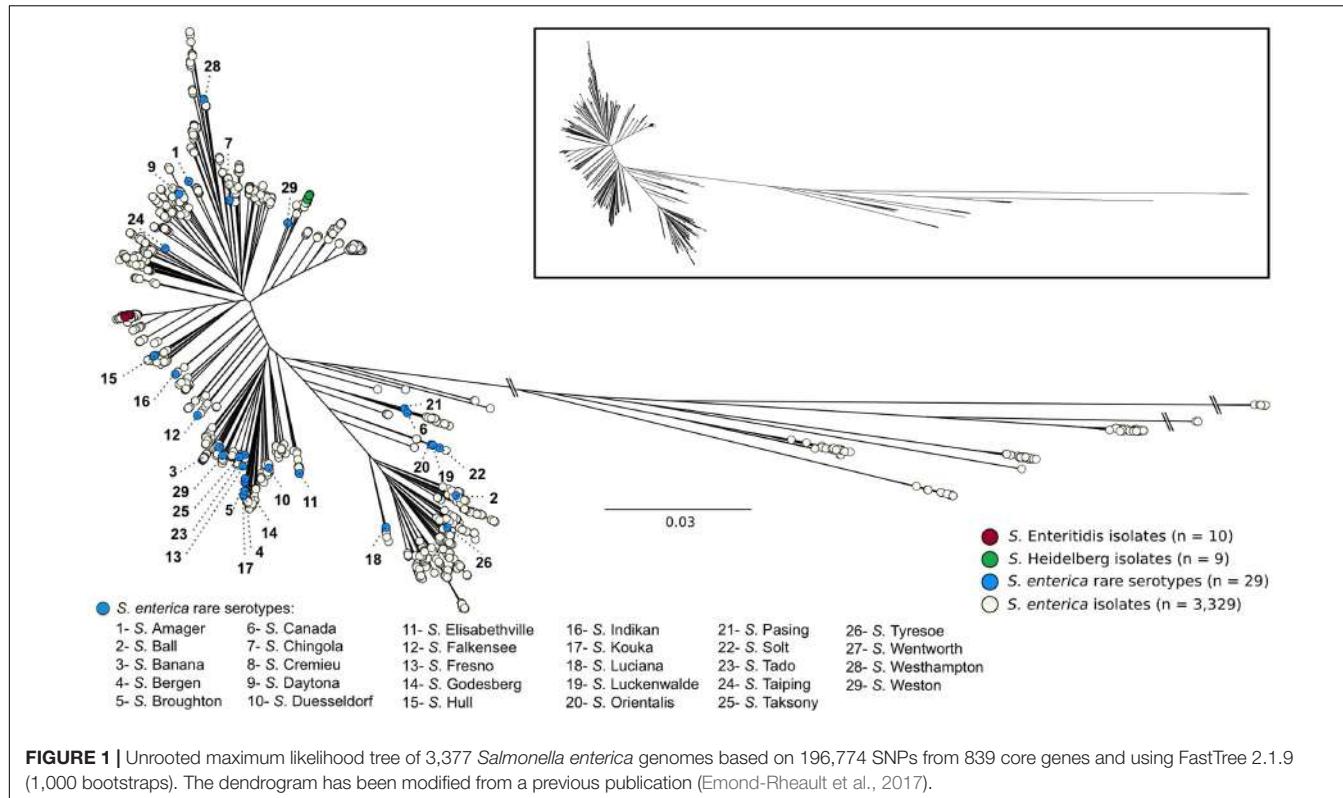
RESULTS

Whole Genome Alignment Reveals Clonality of Clinical *Salmonella* *Enteritidis* and *Salmonella* *Heidelberg* Isolates

Whole genome sequences (WGS) of all *S. enterica* isolates evaluated in this study are contained within the SalFoS database. To investigate the potential for the PCR assay to be used as a rapid screening tool to determine genome diversity in *Salmonella*, we compared the results of the PCR assay with core genome SNP sequence alignments for a population of clonal and diverse *S. enterica* isolates chosen from the SalFoS database. Previously, we sequenced the whole genomes of 3,337 *S. enterica* isolates contained within the SalFoS database, and aligned their core genomes based on SNPs (Emond-Rheault et al., 2017). This data was used to construct an unrooted maximum-likelihood tree of the core genome sequences. Core genome alignment is a subset of whole-genome alignment, which facilitates the construction of large phylogenetic trees between related microorganisms by using the essential genes contained within the core genome (Treangen et al., 2014). Single-nucleotide polymorphisms (SNPs) within the core genome are the most reliable variant to infer large phylogenetic relationships between closely related microorganisms (Treangen et al., 2014). We used core-genome SNP analysis to study the evolution and diversity of *Salmonella*, and observed that *S. enterica* isolates were grouped within two clades, while *Salmonella* from other subspecies clustered within separate clades (Emond-Rheault et al., 2017). These results are in agreement with those of Timme et al. (2013), who studied 156 WGS of *Salmonella* from the six *S. enterica* subspecies by core SNP analysis and observed two clades of *S. enterica* subspecies *enterica*. From our phylogenetic tree, we selected 30 genetically diverse, food associated *Salmonella* isolates from rare serovars (one isolate per serovar), and a clonal population of *Salmonella* consisting of 10 clinical *S. Enteritidis* isolates and 9 clinical *S. Heidelberg* isolates. The food associated *Salmonella* serovars were dispersed throughout the tree (Figure 1) suggesting that these isolates are genetically diverse and could be easily distinguished at the serovar level. In contrast, the *S. Enteritidis* and *S. Heidelberg* isolates were clustered within their corresponding serovar branch with low genetic diversity among strains underscoring the clonal nature of these serovars (Deng et al., 2015; Labbé et al., 2016).

Bioinformatics Analysis Reveals Diversity of Prophages in Rare *Salmonella* Isolates

The prophage finding software PHASTER was used to identify intact prophages and their integrase genes within the food



associated and clinical *Salmonella* isolates. A total of 102 intact prophages and 75 integrase genes (PHASTER was unable to identify an integrase gene in some prophages) were identified by PHASTER. Additionally, PHASTER identified the presence of a large diversity of prophages in the food-associated isolates compared to the clinical isolates (Figure 2). For example, PHASTER identified 29 different types of prophages (and 23 different integrase genes) from nine different bacterial species among *S. enterica* food associated isolates (Figure 2A). The majority of the prophages infect *S. enterica*, while other prophages originate from *Escherichia coli*, *Enterobacteri*a, *Haemophilus influenzae*, *Edwardsiella* spp., *Aeromonas* spp., *Klebsiella* spp., and *Acyrthosiphon spium*. In addition, PHASTER identified two prophages with high homology to *Vibrio* spp. in a *Salmonella* Pasing isolate and a *Salmonella* Elisabethville isolate, and one prophage with high homology to a shiga toxin 2 (Stx2) converting phage in an isolate from serovar Godesberg. In contrast, PHASTER identified only four different types of prophages and four integrase genes within *S. Enteritidis* and *S. Heidelberg* (Figure 2B). The four prophages are the common *Salmonella* phages P22, Gifsy-2, and RE-2010, and the *Haemophilus influenzae* phage HP2.

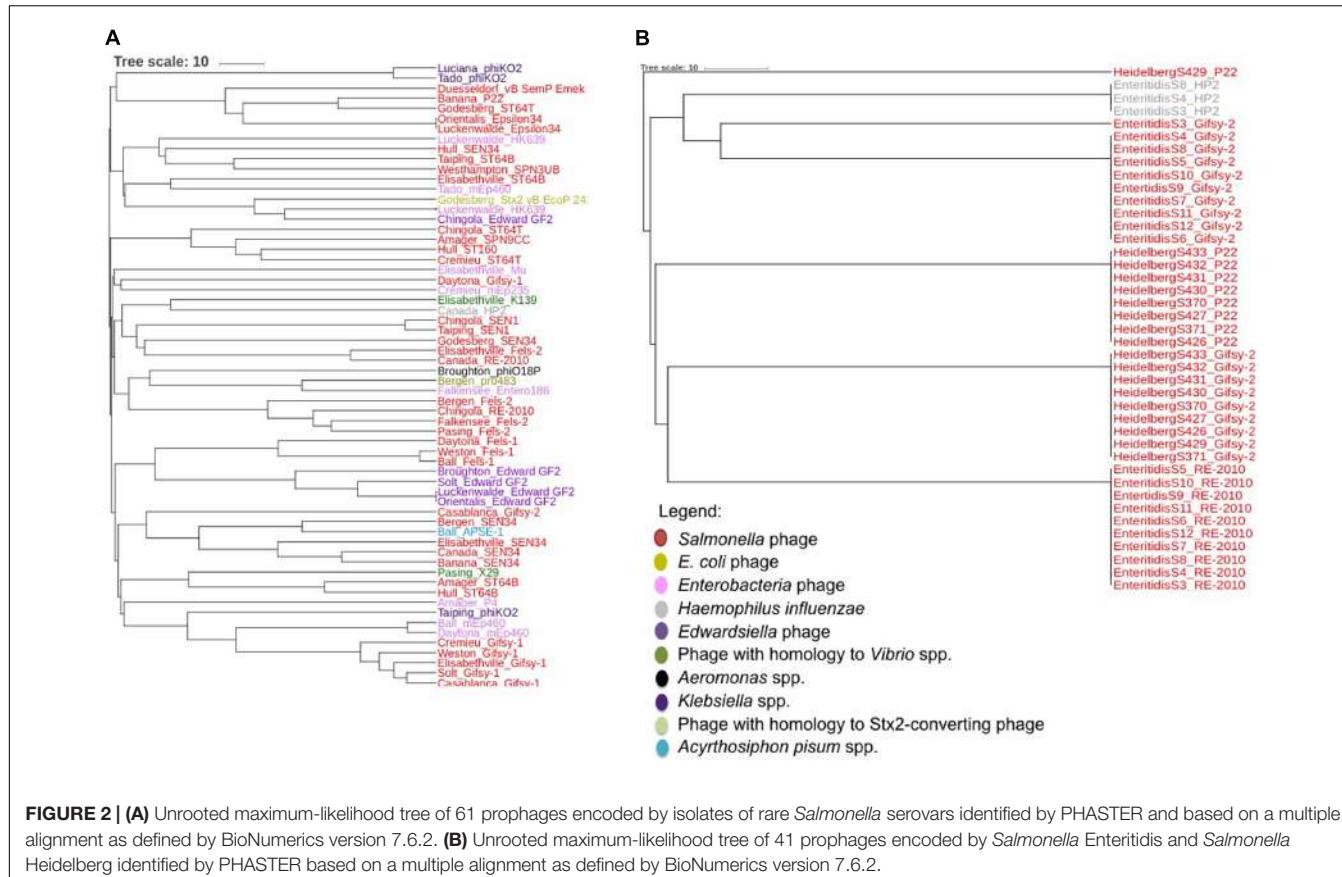
The PCR Assay Reveals Genetic Diversity of Prophage Integrase Genes in Food Associated *Salmonella* Isolates

The PCR assay targets the tyrosine integrase genes of 32 enteric phages infecting members of the *Enterobacteriaceae* family. Using

the PCR assay, a total of 147 integrase genes (79 integrase genes in the food associated *Salmonella* isolates, 50 integrase genes in *S. Enteritidis*, and 18 integrase genes in *S. Heidelberg*) were identified. In agreement with the core genome sequence analysis, the integrase genes from the food associated *Salmonella* isolates were much more diverse than the integrase genes in the clinical *Salmonella* isolates. For example, all 11 primer sets produced amplicons in the food associated isolates, while only five primer sets (1, 4, 6A, 6B, and 7) produced amplicons in the *S. Enteritidis* isolates, and only two primer sets (1 and 7) produced amplicons in the *S. Heidelberg* isolates. Furthermore, all *S. Enteritidis* isolates contained the same prophage integrases, a result that was also observed in the *S. Heidelberg* isolates. Prophage integrase genes amplified by primer sets 2, 3, 5A, 5B, 5C, and 8 were detected in the food associated *Salmonella* isolates, but not the clinical (*S. Enteritidis* and *S. Heidelberg*) isolates (Figure 3).

The Integrase Gene Is an Indicator of Prophage Diversity in *Salmonella*

To further demonstrate that the PCR assay can reveal prophage diversity in *Salmonella* via the integrase gene, a multiple alignment of the integrase gene of the intact prophages identified by PHASTER was performed and used to construct a maximum-likelihood tree (Figure 4). Nine of ten *S. Enteritidis* isolates contained intact Gifsy-2 prophages, and the integrase genes from these phages all clustered together (Figure 4, Cluster 1). Additionally, all *S. Enteritidis* isolates also contained intact RE-2010 prophages, and the integrase genes of these phages



clustered together (**Figure 4**, Cluster 2) in a similar fashion to that observed with the Gifsy-2 integrase genes. These results suggest that the integrase genes from Gifsy-2 and RE-2010 phages are clonal within the isolates of this serovar, and demonstrate that the use of the integrase gene is in agreement with the use of core genome SNP analysis of *S. Enteritidis*, as predictors of genome diversity. The results also indicate that bioinformatic analysis of all prophage integrase genes contained within a given isolate may be used to add discrimination to core genome SNP analysis studies, in cases where the genomes are clonal in nature, as is the case with *S. Enteritidis*. For example, three *S. Enteritidis* isolates (S3, S4, and S8) contain an integrase from the HP2 prophage (**Figure 4**, Cluster 3), which is not found in the other *S. Enteritidis* isolates, and therefore could be used to differentiate between the isolates.

Similar results were observed with the *S. Heidelberg* isolates. For example, all *S. Heidelberg* isolates also contain Gifsy-2 prophages, but the sequence of their integrase genes are different from those contained within the *S. Enteritidis* isolates (**Figure 4**, Cluster 4). Also, while the majority (seven of nine Gifsy-2 integrase genes) formed a cluster, the Gifsy-2 integrase genes from two isolates, *S. Heidelberg* S427 and S430 are outliers and cluster with other similar prophage integrase genes (**Figure 4**). Taken collectively with the *S. Enteritidis* data, these results demonstrate that single integrase genes within a respective isolate can be used to assess genomic diversity, while the

collective number of integrase gene sequences contained within an isolate can be used for discrimination among genetically similar isolates.

DISCUSSION

The study of phage diversity trails far behind similar studies of the bacterial and eukaryotic kingdoms, and only a small fraction of phages have so far been characterized (Adriaenssens and Cowan, 2014). This is largely due to the absence of any universal phylogenetic marker (or signature gene) for phages. Unlike bacteria, in which the 16S rRNA gene is a universal gene that can be used for taxonomy and phylogeny, phages have no universally present gene that can be used for taxonomic analysis (Rohwer and Edwards, 2002). Various signature genes have been used to investigate phage diversity, including genes encoding structural proteins (portal proteins, major capsid proteins, and tail sheath proteins), auxiliary metabolism genes (*psbA*, *psbB*, and *phoH*), and several polymerase genes (Adriaenssens and Cowan, 2014). The majority of work conducted on signature gene analysis to assess phage diversity has focused on virulent phages. With respect to diversity conferred on bacterial hosts by temperate phages, Verghese et al. (2011) investigated the use of comK prophage junction fragments as markers for *Listeria monocytogenes* genotypes that persisted in individual meat and

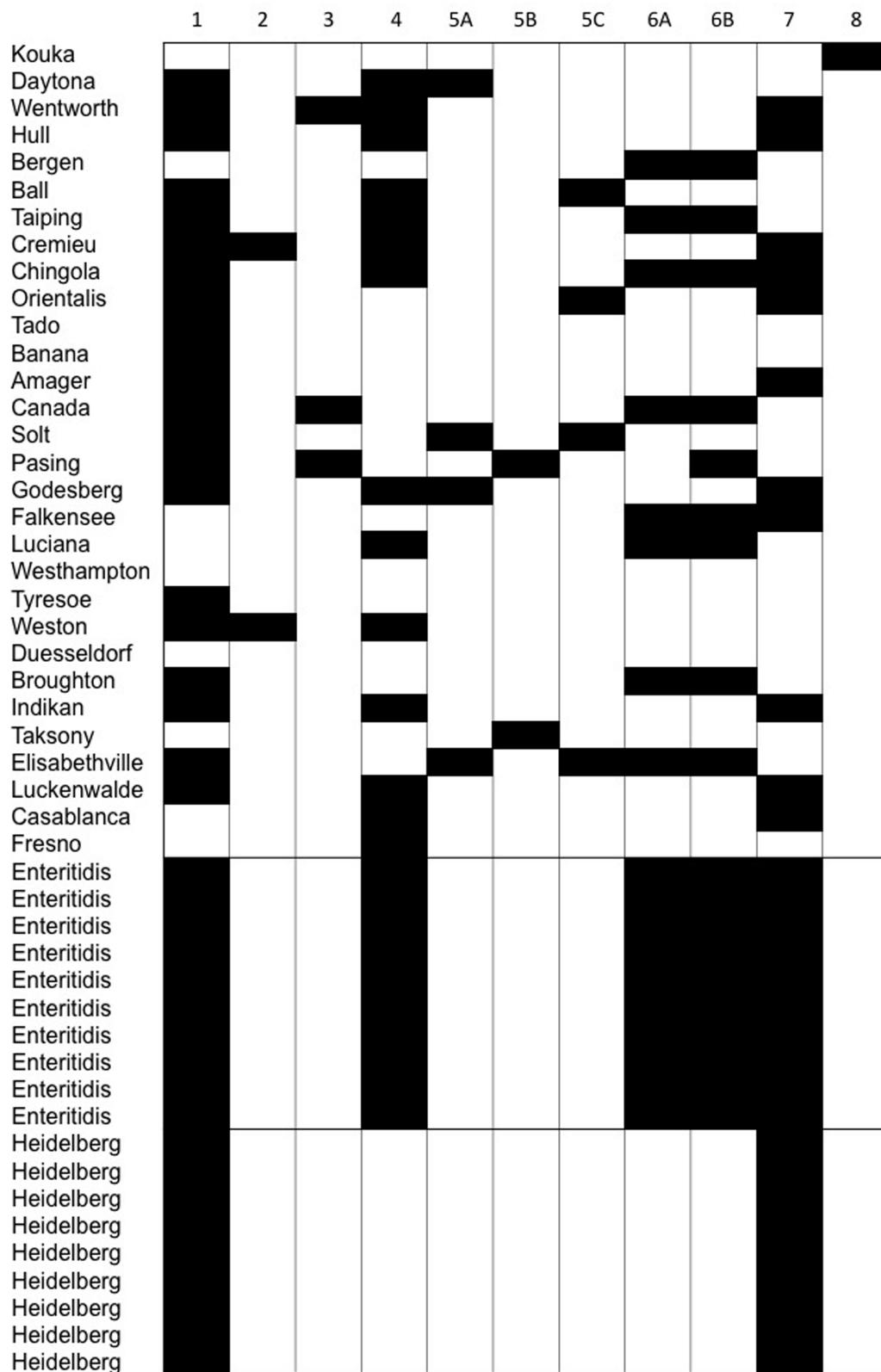


FIGURE 3 | The diversity of phage integrase genes detected by PCR within isolates of rare *Salmonella* serovars compared to the minimal diversity of phages identified within clonal *Salmonella* serovars Enteritidis and Heidelberg.

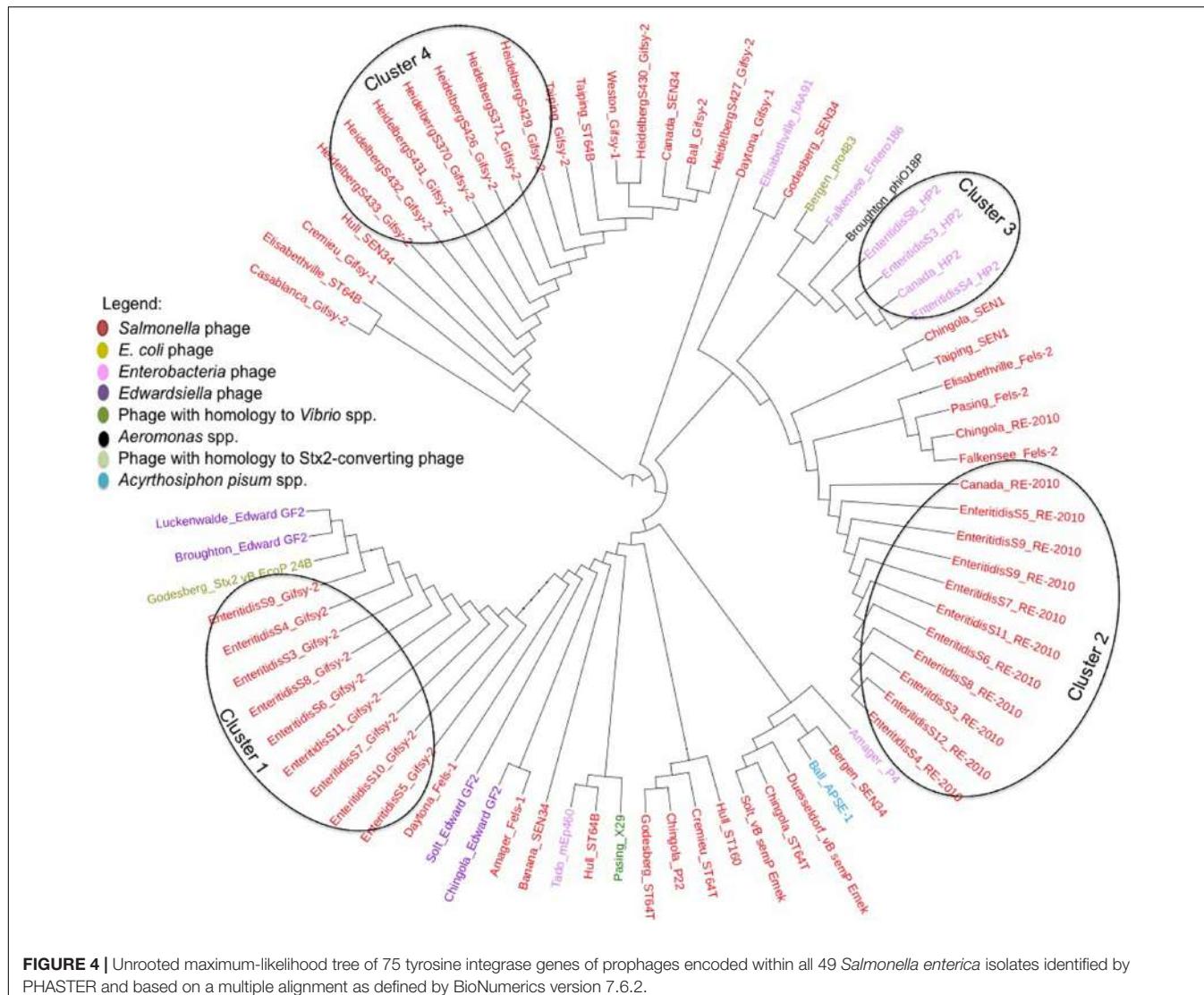


FIGURE 4 | Unrooted maximum-likelihood tree of 75 tyrosine integrase genes of prophages encoded within all 49 *Salmonella enterica* isolates identified by PHASTER and based on a multiple alignment as defined by BioNumerics version 7.6.2.

poultry processing plants (Verghese et al., 2011). In this work, the authors demonstrated that sequences in comK prophage junction fragments could be used to differentiate strains of epidemic clones (ECs), which, when identified, were shown to be specific to individual meat and poultry processing plants. The authors concluded that comK prophage junction fragment sequences may permit accurate tracking of persistent strains within individual food processing operations and thus allow the design of more effective intervention strategies to reduce contamination and enhance food safety (Verghese et al., 2011).

The absence of a universal marker for phages and the fact that signature genes are only specific for certain phages or phage species, limits the use of such approaches to related phages, meaning that these approaches are not useful when assessing phages that are genetically diverse. Hence, in this work, we used the prophage integrase gene as a signature gene to measure prophage diversity, and therefore the diversity of host genomes that carry the prophages. Many factors such as pathogenicity

islands, virulence factors, and antimicrobial and heavy metal resistance genes are encoded by prophages and provide genetic diversity among bacterial species (Brüssow et al., 2004). We therefore hypothesized that prophages may contribute to the diversity in the *Salmonella* isolates from rare serovars, and a lack of diversity in the clonal (clinical) serovars. The tyrosine integrase gene can be used as an indicator for phage diversity because it is carried by over 300 prophages (Balding et al., 2005). Although prophages can carry serine integrase genes, only approximately 30 members have been identified and they are not typically carried by foodborne bacteria (Fogg et al., 2014). The C-terminal of the tyrosine integrase contains two highly conserved regions designated “box I” and “box II.” The tyrosine residue is located at residue 342 in “box II” and is conserved in every family member encoding a tyrosine integrase and is responsible for the DNA cleavage on the bacterial attachment site (attB) and the phage attachment site (attP), which facilitate the integration of the prophage into its host (Balding et al., 2005).

In addition, two arginine residues R212 and R311 that assist in the facilitation of prophage integration are also conserved in all family members. Bobay et al. (2013) investigated the number and type of integration loci of prophages within *E. coli* and *S. enterica* genomes. They observed that the most prevalent (46%) integration site flanking prophages in *S. enterica* was a protein coding sequence, *lepA*, followed by tRNA and tmRNA genes (37%) and sRNA genes (17%). Of 24 prophage integration sites observed within *S. enterica*, 19 were also present in *E. coli*, suggesting that prophage integration sites are restricted to a few bacterial sites. The authors also observed that 423 of 500 prophages (83%) contained an integrase, of which all were tyrosine integrases, and a phylogenetic analysis demonstrated that closely related integrases integrate at the same integration sites. Since phages restrict their integration to conserved sites on the host genome, this further supports our findings that the integrase gene can serve as an indicator of prophage diversity (Bobay et al., 2013).

Core genome SNP analysis of the *Salmonella* isolates demonstrated a high degree of clonality in the *S. Enteritidis* and *S. Heidelberg* isolates (**Figure 1**), in agreement with previous studies, while also showing that the food associated *Salmonella* were highly diverse. Ogunremi et al. (2014) observed the clonality of 11 *S. Enteritidis* isolates, and demonstrated that they were all characterized by a comparably sized genome, an estimated 23–905 SNPs among genome pairs and five prophages or prophage remnants. Hoffmann et al. (2014) observed the clonality of 44 *S. Heidelberg* isolates, in which nearly 30 were indistinguishable by pulsed-field gel electrophoresis (PFGE), had significantly fewer core genome SNPs than *S. Newport* and, *S. Typhimurium* and contained similar prophages.

In this study, the results of the core genome SNP analysis agreed with a prophage integrase gene bioinformatic approach, in which prophages and integrase genes were identified within WGS using PHASTER, and used to construct unrooted maximum likelihood trees. The integrase gene was a good predictor of the diversity of the entire prophage sequence, as prophages that clustered together (i.e., Gifsy-2, HP2, and RE-2010) (**Figure 2**), contained integrase genes that also clustered together (**Figure 4**). Gifsy-2 is a highly studied phage known to integrate into various

strains of *S. Typhimurium*. HP2 is a P2-like phage known to infect *Salmonella* spp., while RE-2010 has high nucleotide homology to Fels-2 prophages that have been observed in various *Salmonella* genomes (Switt et al., 2015). Also, both the entire prophage sequences and the integrase gene sequences were good predictors of *Salmonella* core genome diversity, as only four intact prophages and their integrase genes were identified in the clonal *Salmonella* (*Enteritidis* and *Heidelberg*) serovars, while 29 different types of prophages and 23 different integrase genes, were identified in the food associated *Salmonella* isolates. These results agreed with the core genome SNP analysis of the *Salmonella* isolates, in which similar results regarding the genome diversity of the clinical and food associated isolates were observed.

As with the bioinformatics approach, the PCR assay detected more integrase genes in the food associated isolates than the clinical isolates. Additionally, when the PCR assay was used to evaluate the *Salmonella* isolates for the presence of integrase genes, the results showed that the PCR assay detected 147 integrases in the 49 *Salmonella* isolates, while 74 integrases were detected by PHASTER in the intact prophages encoded by the *Salmonella* isolates (**Figure 5**). The differential ability of the two methods to detect prophage integrase genes is due to the fact that the primer sets used in the PCR assay detected the integrase genes of intact and cryptic phages, while our bioinformatic analysis focused only on the intact phages detected by PHASTER. The prophages encoded by the *Salmonella* isolates were identified as intact, questionable or incomplete by PHASTER. Incomplete phages suggest that they may represent cryptic phages, which may no longer encode an integrase gene. Questionable phages do not contain sufficient prophage genes to be considered complete functional phages. Thus, only intact phages were selected and blasted for confirmation to ensure an accurate multiple sequence alignment.

Other groups have demonstrated the use of prophages as markers of diversity in their bacterial hosts. For example, (Shan et al., 2012) conducted phylogenetic analysis of the holin sequences of *Clostridium difficile* prophages, and identified three groups of *C. difficile* phages, two within the Myoviridae and a divergent group within the Siphoviridae. The marker also produced homogenous groups within temperate phages that

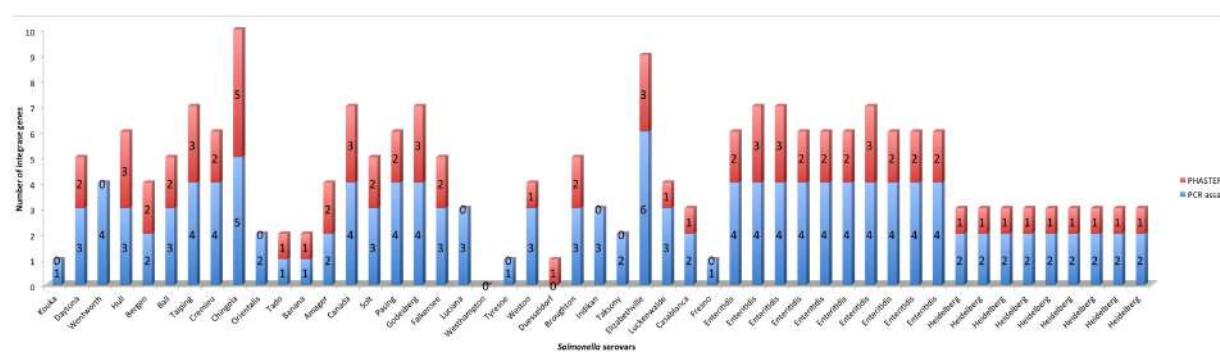


FIGURE 5 | Comparison of the number of integrase genes detected by the PCR assay to the number of integrase genes detected by PHASTER within all 49 *Salmonella enterica* genomes.

infect other taxa, including *Clostridium perfringens*, *Clostridium botulinum*, and *Bacillus* spp., indicating the potential use of the holin gene to study prophage carriage in other bacteria. This study demonstrated the high incidence of prophage carriage and diversity in clinically relevant strains of *C. difficile*.

Basu et al. (2000) conducted a molecular analysis of CTX prophages in clinical, classical biotype strains of *Vibrio cholerae* O1 that were isolated between 1970 and 1979 (Basu et al., 2000). Restriction fragment length polymorphism (RFLP) of rRNA genes and PFGE showed clonal diversity among the strains. The authors observed that one strain (GP13) had three CTX prophages while another (GP147) had four CTX prophages, indicating heterogeneity in the arrangement of the CTX prophages among classical strains of *V. cholerae* O1.

Collectively, our data shows that the prophage integrase PCR assay may be a good indicator of genome diversity in *S. enterica*, and that the PCR assay is a rapid and cost-effective rapid screening tool that may be used as a high throughput screen to evaluate large numbers of *Salmonella* isolates as a way to reduce

the numbers of isolates that are submitted for whole genome sequencing to evaluate genomic diversity.

AUTHOR CONTRIBUTIONS

SB and ST supplied bacterial isolates and AC, YD, ET, LG, J-GE-R, JJ, LF, IK-I, and RL performed the analyses and drafted the manuscript. BB provided support for sequencing and analysis. All authors revised the manuscript.

ACKNOWLEDGMENTS

We express our gratitude to members of the genomics analysis and bioinformatics platforms at IBIS. LG, RL, SB, and ST are funded by Genome Canada and by the Génome Québec provincial genome center. LG is funded by the National Sciences and Engineering Council of Canada Discovery Grants Program (grant number RGPIN-2014-0574).

REFERENCES

- Adriaenssens, E. M., and Cowan, D. A. (2014). Using signature genes as tools to assess environmental viral ecology and diversity. *Appl. Environ. Microbiol.* 80, 4470–4480. doi: 10.1128/AEM.00878-14
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2
- Applied Maths (2017). *BioNumerics Version 7.6.2*. Available at: <http://www.applied-maths.com>
- Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi: 10.1093/nar/gkw387
- Balding, C., Bromley, S. A., Pickup, R. W., and Saunders, J. R. (2005). Diversity of phage integrases in *Enterobacteriaceae*: development of markers for environmental analysis of temperate phages. *Environ. Microbiol.* 7, 1558–1567. doi: 10.1111/j.1462-2920.2005.00845.x
- Basu, A., Mukhopadhyay, A. K., Garg, P., Chakraborty, S., Ramamurthy, T., Yamasaki, S., et al. (2000). Diversity in the arrangement of the CTX prophages in classical strains of *Vibrio cholerae* O1. *FEMS Microbiol. Lett.* 182, 35–40. doi: 10.1111/j.1574-6968.2000.tb08869.x
- Bobay, L.-M., Rocha, E. P., and Touchon, M. (2013). The adaptation of temperate bacteriophages to their host genomes. *Mol. Biol. Evol.* 30, 737–751. doi: 10.1093/molbev/mss279
- Brüssow, H., Canchaya, C., and Hardt, W.-D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* 68, 560–602. doi: 10.1128/MMBR.68.3.560-602.2004
- Callejón, R. M., Rodríguez-Naranjo, M. I., Ubeda, C., Hornero-Ortega, R., García-Parrilla, M. C., and Troncoso, A. M. (2015). Reported foodborne outbreaks due to fresh produce in the United States and European Union: trends and causes. *Foodborne Pathog. Dis.* 12, 32–38. doi: 10.1089/fpd.2014.1821
- Deng, X., Shariat, N., Driebe, E. M., Roe, C. C., Tolar, B., Trees, E., et al. (2015). Comparative analysis of subtyping methods against a whole-genome sequencing standard for *Salmonella enterica* serotype Enteritidis. *J. Clin. Microbiol.* 53, 212–218. doi: 10.1128/JCM.02332-14
- Emond-Rheault, J.-G., Jeukens, J., Freschi, L., Kukavica-Ibrulj, I., Boyle, B., Dupont, M.-J., et al. (2017). A Syst-OMICS approach to ensuring food safety and reducing the economic burden of salmonellosis. *Front. Microbiol.* 8:996. doi: 10.3389/fmicb.2017.00996
- Figueroa-Bossi, N., and Bossi, L. (1999). Inducible prophages contribute to *Salmonella* virulence in mice. *Mol. Microbiol.* 33, 167–176. doi: 10.1046/j.1365-2958.1999.01461.x
- Fogg, P. C., Colloms, S., Rosser, S., Stark, M., and Smith, M. C. (2014). New applications for phage integrases. *J. Mol. Biol.* 426, 2703–2716. doi: 10.1016/j.jmb.2014.05.014
- Hoffmann, M., Zhao, S., Pettengill, J., Luo, Y., Monday, S. R., Abbott, J., et al. (2014). Comparative genomic analysis and virulence differences in closely related *Salmonella enterica* serotype Heidelberg isolates from humans, retail meats, and animals. *Genome Biol. Evol.* 6, 1046–1068. doi: 10.1093/gbe/evu079
- Kropinski, A. M., Sulakvelidze, A., Koncny, P., and Poppe, C. (2007). *Salmonella* phages and prophages—genomics and practical aspects. *Methods Protoc.* 394, 133–175. doi: 10.1007/978-1-59745-512-1_9
- Labbé, G., Ziebell, K., Bekal, S., Macdonald, K. A., Parmley, E. J., Agunos, A., et al. (2016). Complete genome sequences of 17 Canadian isolates of *Salmonella enterica* subsp. *enterica* serovar Heidelberg from human, animal, and food sources. *Genome Announc.* 4:e00990-16. doi: 10.1128/genomeA.00990-16
- Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242–W245. doi: 10.1093/nar/gkw290
- Majowicz, S., Musto, J., Scallan, E., Angulo, F., Kirk, M., O'Brien, S., et al. (2010). The global burden of nontyphoidal *Salmonella* gastroenteritis. *Clin. Infect. Dis.* 50, 882–889. doi: 10.1086/650733
- McNair, K., Bailey, B. A., and Edwards, R. A. (2012). PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* 28, 614–618. doi: 10.1093/bioinformatics/bts014
- Ogunremi, D., Devenish, J., Amoako, K., Kelly, H., Dupras, A. A., Belanger, S., et al. (2014). High resolution assembly and characterization of genomes of Canadian isolates of *Salmonella* Enteritidis. *BMC Genomics* 15:713. doi: 10.1186/1471-2164-15-713
- Public Health Agency of Canada [PHAC] (2012). *Executive Summary for the National Enteric Surveillance Program 2012 Annual Report*. Available at: http://publications.gc.ca/collections/collection_2014/aspc-phac/HP37-15-2012-eng.pdf [accessed May 22, 2017].
- Rohwer, F. (2003). Global phage diversity. *Cell* 113, 141. doi: 10.1016/S0092-8674(03)00276-9
- Rohwer, F., and Edwards, R. (2002). The phage proteomic tree: a genome-based taxonomy for phage. *J. Bacteriol.* 184, 4529–4535. doi: 10.1128/JB.184.16.4529-4535.2002
- Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M.-A., Roy, S. L., et al. (2011). Foodborne illness acquired in the United States—major pathogens. *Emerg. Infect. Dis.* 17, 7–15. doi: 10.3201/eid1701.P11101
- Shan, J., Patel, K. V., Hickenbotham, P. T., Nale, J. Y., Hargreaves, K. R., and Clokie, M. R. (2012). Prophage carriage and diversity within clinically

- relevant strains of *Clostridium difficile*. *Appl. Environ. Microbiol.* 78, 6027–6034. doi: 10.1128/AEM.01311-12
- Switt, A. I. M., Sulakvelidze, A., Wiedmann, M., Kropinski, A. M., Wishart, D. S., Poppe, C., et al. (2015). *Salmonella* phages and prophages: genomics, taxonomy, and applied aspects. *Methods Protoc.* 1225, 237–287. doi: 10.1007/978-1-4939-1625-2_15
- Thomas, M. K., Murray, R., Flockhart, L., Pintar, K., Fazil, A., Nesbitt, A., et al. (2015). Estimates of foodborne illness-related hospitalizations and deaths in Canada for 30 specified pathogens and unspecified agents. *Foodborne Pathog. Dis.* 12, 820–827. doi: 10.1089/fpd.2015.1966
- Thomas, M. K., Murray, R., Flockhart, L., Pintar, K., Pollari, F., Fazil, A., et al. (2013). Estimates of the burden of foodborne illness in Canada for 30 specified pathogens and unspecified agents, circa 2006. *Foodborne Pathog. Dis.* 10, 639–648. doi: 10.1089/fpd.2012.1389
- Thomson, N., Baker, S., Pickard, D., Fookes, M., Anjum, M., Hamlin, N., et al. (2004). The role of prophage-like elements in the diversity of *Salmonella enterica* serovars. *J. Mol. Biol.* 339, 279–300. doi: 10.1016/j.jmb.2004.03.058
- Timme, R. E., Pettengill, J. B., Allard, M. W., Strain, E., Barrangou, R., Wehnert, C., et al. (2013). Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. *Genome Biol. Evol.* 5, 2109–2123. doi: 10.1093/gbe/evt159
- Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15, 524. doi: 10.1186/s13059-014-0524-x
- Tritt, A., Eisen, J. A., Facciotti, M. T., and Darling, A. E. (2012). An integrated pipeline for de novo assembly of microbial genomes. *PLoS ONE* 7:e42304. doi: 10.1371/journal.pone.0042304
- Verghese, B., Lok, M., Wen, J., Alessandria, V., Chen, Y., Kathariou, S., et al. (2011). comK prophage junction fragments as markers for *Listeria monocytogenes* genotypes unique to individual meat and poultry processing plants and a model for rapid niche-specific adaptation, biofilm formation, and persistence. *Appl. Environ. Microbiol.* 77, 3279–3292. doi: 10.1128/AEM.00546-11
- Wain, J., and O'Grady, J. (2017). "Genomic diversity in *Salmonella enterica*," in *Applied Genomics of Foodborne Pathogens*, eds X. Deng, H. C. den Bakker, and R. S. Hendriksen (Cham: Springer), 91–107.
- Zhou, Z., Mccann, A., Litrap, E., Murphy, R., Cormican, M., Fanning, S., et al. (2013). Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* serovar Agona. *PLoS Genet.* 9:e1003471. doi: 10.1371/journal.pgen.1003471

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor declared a shared affiliation, though no other collaboration, with one of the authors ST, and the handling Editor states that the process met the standards of a fair and objective review.

Copyright © 2017 Colavecchio, D'Souza, Tompkins, Jeukens, Freschi, Emond-Rheault, Kukavica-Ibrulj, Boyle, Bekal, Tamer, Levesque and Goodridge. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Temporal Genomic Phylogeny Reconstruction Indicates a Geospatial Transmission Path of *Salmonella Cerro* in the United States and a Clade-Specific Loss of Hydrogen Sulfide Production

Jasna Kovac¹, Kevin J. Cummings², Lorraine D. Rodriguez-Rivera², Laura M. Carroll¹, Anil Thachil³ and Martin Wiedmann^{1*}

¹ Department of Food Science, Cornell University, Ithaca, NY, USA, ² Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, USA, ³ Department of Population Medicine and Diagnostic Sciences, Cornell University, Ithaca, NY, USA

OPEN ACCESS

Edited by:

Jennifer Ronholm,
McGill University, Canada

Reviewed by:

David Rodriguez-Lazaro,
University of Burgos, Spain
Beatrix Stessl,
Veterinärmedizinische Universität,
Austria

*Correspondence:

Martin Wiedmann
mw16@cornell.edu

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 20 February 2017

Accepted: 10 April 2017

Published: 01 May 2017

Citation:

Kovac J, Cummings KJ, Rodriguez-Rivera LD, Carroll LM, Thachil A and Wiedmann M (2017)
Temporal Genomic Phylogeny Reconstruction Indicates a Geospatial Transmission Path of *Salmonella Cerro* in the United States and a Clade-Specific Loss of Hydrogen Sulfide Production.
Front. Microbiol. 8:737.
doi: 10.3389/fmicb.2017.00737

Salmonella Cerro has become one of the most prevalent *Salmonella* serotypes isolated from dairy cattle in several U.S. states, including New York where it represented 36% of all *Salmonella* isolates of bovine origin in 2015. This serotype is commonly isolated from dairy cattle with clinical signs of salmonellosis, including diarrhea and fever, although it has also been identified in herds without evidence of clinical disease or decreased production. To better understand the transmission patterns and drivers of its geographic spread, we have studied the genomic similarity and microevolution of *S. Cerro* isolates from the northeast U.S. and Texas. Eighty-three out of 86 isolates were confirmed as multilocus sequence type 367. We identified core genome SNPs in 57 upstate New York (NY), 2 Pennsylvania (PA), and 27 Texas *S. Cerro* isolates from dairy cattle, farm environments, raw milk, and one human clinical case and used them to construct a tip-dated phylogeny. *S. Cerro* isolates clustered in three distinct clades, including (i) clade I ($n = 3$; 2013) comprising isolates from northwest Texas (NTX), (ii) clade II ($n = 14$; 2009–2011, 2014) comprising isolates from NY, and (iii) clade III comprising isolates from NY, PA, and central Texas (CTX) in subclade IIIa ($n = 45$; 2008–2014), and only CTX isolates in subclade IIIb ($n = 24$; 2013). Temporal phylogenetic analysis estimated the divergence of these three clades from the most recent common ancestor in approximately 1980. The CTX clade IIIb was estimated to have evolved and diverged from the NY ancestor around 2004. Furthermore, gradual temporal loss of genes encoding a D-alanine transporter, involved in virulence, was observed. These genes were present in the isolates endemic to NTX clade I and were gradually lost in clades II and III. The virulence gene *orgA*, which is part of the *Salmonella* Pathogenicity Island 1, was lost in a subgroup of Texas isolates in clades I and IIIb. All *S. Cerro* isolates had an additional cytosine inserted in a cytosine-rich region of the virulence gene *sopA*, resulting in premature termination of translation likely responsible for loss of pathogenic

capacity in humans. A group of closely related NY isolates was characterized by the loss of hydrogen sulfide production due to the truncation or complete loss of *phsA*. Our data suggest the ability of *Salmonella* to rapidly diverge and adapt to specific niches (e.g., bovine niche), and to modify virulence-related characteristics such as the ability to utilize tetrathionate as an alternative electron acceptor, which is commonly used to detect *Salmonella*. Overall, our results show that clinical outcome data and genetic data for *S. Cerro* isolates, such as truncations in virulence genes leading to novel pheno- and pathotypes, should be correlated to allow for accurate risk assessment.

Keywords: *Salmonella enterica* subsp. *enterica* serotype Cerro, dairy, WGS, emerging pathogen, epidemiology, virulence genes, hydrogen sulfide

INTRODUCTION

Dairy cattle represent a reservoir of a number of *Salmonella* serotypes. Some of these, such as Typhimurium, 4,5,12:i:-, Newport, and Montevideo, are frequently implicated in human infections, while others, such as Cerro, rarely cause disease in humans (Jones et al., 2008; Rodriguez-Rivera et al., 2016). This latter serotype is of concern because of frequent association with clinical disease in cattle (diarrhea, fever, depression, and decreased appetite), which can result in increased treatment and labor expenses, reduced milk yield, and loss of animals through mortality or culling (United States Department of Agriculture [USDA], 2011). Furthermore, this serotype will generate a positive result in *Salmonella* detection assays and will be considered as an adulterant in food, despite the low risk for human infection. Such outcomes are consequently associated with increased economic burden for farmers and industry.

Salmonella Cerro has been considered an emerging pathogen in past years due to its increased prevalence among dairy cattle (Cummings et al., 2010a,b; Tewari et al., 2012). High prevalence of *S. Cerro* was reported among dairy herds (20/57 herds; 35%) in NY between 2007 and 2009 (Cummings et al., 2010b). *S. Cerro* was isolated from 59% of the dairy cattle with clinical evidence of salmonellosis in that study (Cummings et al., 2010b). Nevertheless, it has also remained one of the most common serotypes recovered from dairy cattle without clinical signs (Rodriguez-Rivera et al., 2014b). One of the first *S. Cerro* subclinical outbreaks was documented in PA between 2004 and 2006 (Van Kessel et al., 2007), which was not an isolated case, as *S. Cerro* was detected in several other PA farms in the region (Van Kessel et al., 2013). The proportional prevalence of *S. Cerro* among *Salmonella* positive cases in PA has approximately doubled between 2005 and 2010 (from 14.3 to 36.1%) (Tewari et al., 2012), demonstrating its rapid emergence in this state.

More frequent isolation of *S. Cerro* was also recently reported in the U.S. Midwest (Hong et al., 2016; Valenzuela et al., 2017). The proportional prevalence of *S. Cerro* among *Salmonella*

gradually increased from less than 1% (2006) to 37% (2015) in Wisconsin, where it became a predominant bovine-associated serotype in 2013 (Valenzuela et al., 2017). A similar trend was observed in Minnesota, where *S. Cerro* accounted for 6.6% ($n = 68$) of all *Salmonella* isolated from cattle and ranked third among most common serotypes from bovine sources between 2006 and 2015 (Hong et al., 2016). Recently, *Salmonella* was also recovered from 67% ($n = 236$) of environmental samples and 64% ($n = 43$) of bovine fecal samples from 11 dairy farms in Texas (Rodriguez-Rivera et al., 2016); serotype Cerro was identified on several of these farms (unpublished data), further suggesting that this serotype is emerging across the country.

The successful spread of *S. Cerro* is likely supported by its adaptation to the bovine host, as suggested by the persistent, estimated 7-month long mean duration of infection (Chapagain et al., 2008). One of the contributing factors is also a high basic reproduction number ($R_0 = 5.8$) (Chapagain et al., 2008), which indicates a rapid spread in dairy cattle. Consequently, *S. Cerro* remains challenging to control at the farm level. While *S. Cerro* commonly causes disease in cattle, it is rarely implicated in clinical human infections (Tewari et al., 2012) and therefore represents a good model system for studying virulence of *Salmonella*. Genomic analysis has previously revealed a gradual loss of D-alanine transporter and a mutation in *sopA* virulence gene that resulted in truncation, which likely influenced decreased virulence of *S. Cerro* in humans (Rodriguez-Rivera et al., 2014a). Hydrogen sulfide, a product of thiosulfate respiration, has been shown to provide a competitive advantage to *Salmonella* in human hosts and is also considered a virulence factor (Winter et al., 2010). Hydrogen sulfide-producing *Salmonella* colonies appear black on selective differential agars used in standard microbiological isolation protocols (Food and Drug Administration); therefore, this characteristic also plays an important role in successful detection and identification of this pathogen. Emergence of *Salmonella* isolates with an impaired ability to produce hydrogen sulfide is concerning, as this phenotype increases the risk for false negative detection of *Salmonella* using traditional microbiological methods, which rely on characteristic black color of hydrogen sulfide precipitate on selective differential media, such as XLD, HE, and BS agars recommended by the Food and Drug Administration Bacteriological Analytical Manual (Food and Drug Administration).

Abbreviations: CTX, central Texas; ESS, effective sample size; GTR, general time-reversible substitution model; HDP, highest probability density; ML, maximum likelihood; MLST, multilocus sequence typing; NTX, northwest Texas; NY, New York; PA, Pennsylvania; PCA, principal component analysis; U.S., United States; WGS, whole genome sequence.

In the present study, we analyzed 86 genomic sequences of *S. Cerro* isolated between 2008 and 2014 in NY, PA, and TX to (i) better understand geographical and temporal spread of serotype Cerro in the U.S., (ii) identify geospatial accumulation of genomic changes potentially linked with virulence, and (iii) examine the ability of these isolates to produce hydrogen sulfide.

MATERIALS AND METHODS

Isolate Selection

New York and PA isolates were selected from a pool of 1,645 *Salmonella enterica* subsp. *enterica* serotype Cerro isolates deposited in a Food Microbe Tracker database at Cornell University (Vangay et al., 2013), to represent a wide range of years (2008–2014) and sources (animal, environment, food, human). One human *S. Cerro* isolate (FSL R8-4516; isolate from a stool sample of a human sporadic case from September 2009) obtained from New York State Department of Health was included in the study for comparison of virulence profiles with bovine isolates. Texas isolates were randomly selected from a pool of *S. Cerro* isolates obtained through a recent field study (Rodriguez-Rivera et al., 2016), using a random number generator¹. These isolates were obtained from Texas dairy farm environments and cull cow fecal samples.

DNA Extraction and Whole Genome Sequencing

The 86 *S. Cerro* isolates from upstate NY ($n = 54$), PA ($n = 2$), CTX ($n = 27$), and NTX ($n = 3$) were whole genome sequenced and analyzed. Frozen cultures (-80°C) in 15% v/v glycerol-BHI media were streaked on BHI agar and incubated for 24 h at 32°C . The DNA of isolates was extracted using the QIAamp DNA MiniKit (Qiagen, Valencia, CA, USA), following the manufacturer's protocol. DNA was eluted in 50 μl Tris-HCl (pH 8.0), and double-stranded DNA (dsDNA) was quantified with Picogreen (Invitrogen, Paisley, UK). DNA that was used for construction of Nextera XT libraries (Illumina, Inc., San Diego, CA, USA) was normalized to a concentration of 0.2 ng/ μl , and sequenced on an Illumina MiSeq platform with 250 bp paired end reads (Genomics Facility of the Cornell University Institute of Biotechnology) (Supplementary Table S1). DNA that was used for construction of NextFlex libraries (Bioo Scientific, Austin, TX, USA) was normalized to concentration of 6.7 ng/ μl and sequenced on an Illumina HiSeq 2500 platform with 250 bp paired end reads (Texas A&M Genomics and Bioinformatics Service). Sequences were analyzed following the workflow described in the next paragraphs and in the Supplementary Material file log.sh; scripts were deposited on GitHub².

Whole Genome Sequence Quality Control, Assembly, Annotation and MLST

Sequencing adapters were trimmed and low quality bases removed with Trimmomatic 0.33 (Bolger et al., 2014) using

default settings, and Nextera XT PE or NextFlex PE adapter sequence files (Supplementary Material file log.sh). Quality of trimmed reads was checked using FastQC v0.11.2 (Babraham Bioinformatics, 2014) prior to *de novo* assembly with SPAdes 3.6.0 (Bankevich et al., 2012). Quality of draft genomes was evaluated using QUAST 3.2, and average coverage computed using BBmap 35.49 (Bushnell, 2015) and Samtools 1.3.1 (Li et al., 2009). Draft genomes were annotated through RASTtk (Brettin et al., 2015). MLST were determined using SRST2 (Inouye et al., 2014).

Core Genome Phylogeny

Single nucleotide polymorphisms (SNPs) were identified by kSNP v2 in 86 *S. Cerro* draft genomes using an optimal kmer size of 19, as determined by Kchooser (Gardner and Hall, 2013). Identified SNPs were used to construct initial ML phylogeny with general time-reversible (GTR) model and 1000 bootstrap iterations in RaxML version 8 (Stamatakis, 2014). This initial tree and assembly quality metrics were used to guide the selection of a reference strain FSL R8-3655 for comprehensive variant calling using cortex_var (Iqbal et al., 2012). Cortex_var was run with the kmer sizes of 33 and 63 to identify variants across 86 *S. Cerro* genomes, using strain FSL R8-3655 as a reference; high quality SNPs (qtresh set at 15) were used in further analyses. Gubbins 1.4.2 (Croucher et al., 2015) was used to identify potential regions of recombination that needed to be filtered out prior to Bayesian phylogenetic inference.

Molecular clock hypothesis of all tips of the tree being equidistant from the root of the tree was tested using SNP tree topology and sequence alignment in MEGA 6.06 (Tamura et al., 2013). The molecular clock hypothesis was further evaluated in MEGA with Tajima's relative rate test based on three representatives of different lineages (i.e., BOV1-0254 representing clade I, FSL R8-3460 representing clade II, and BOV1-0002 representing clade IIIb). Significance of differences between the log-likelihoods obtained with and without the molecular clock assumption was calculated using chi-squared statistics in R.

Linear regression models implemented in TempEst v1.5 (Rambaut et al., 2016) were used to evaluate the temporal signal and clocklikeness of the phylogeny based on associations between temporal sequence divergence and isolation dates. Statistical significance of the obtained correlation coefficient was assessed using t-statistics in R.

A tip-dated phylogeny was constructed using BEAUTi v1.8.2 and BEAST v1.8.2 (Drummond et al., 2012) with a combination of the GTR substitution model and (i) strict clock and coalescent constant size population models, (ii) strict clock and coalescent Bayesian skyline models, (iii) lognormal relaxed clock and coalescent constant size population models, and (iv) lognormal relaxed clock and coalescent Bayesian skyline models. The initial runs (seed 123456) were carried out with a substitution rate prior set to 2.4×10^{-7} /site/year. This substitution rate was estimated in a previous study of *S. Cerro* evolution (Rodriguez-Rivera et al., 2014a). An ascertainment bias correction was used to account for the use of solely variant sites (Supplementary Material file log.sh). Markov Chain Monte Carlo (MCMC) algorithm was run for

¹random.org/random.org

²https://github.com/jasnakovac/salmonella_cerro

100 million generations, and parameters were logged every 1000 generations. Marginal likelihood estimations were computed by path sampling in 100 steps with a chain length of 1,000,000 and likelihoods logged every 1,000 generations. The best model combination was identified based on a combination of (i) the mean marginal likelihood values from these two runs and (ii) ESS of run statistics (e.g., prior, posterior, tree likelihood, clock rate, and coalescent). The best model combination was used to run three additional 100,000,000 MCMC runs with different random seeds (i.e., 654321, 2739 and 098765), and priors used in the first run. The output statistics and traces were analyzed in Tracer v1.6.0, and the log and trees files of converging individual runs were combined in LogCombiner v1.8.3 (burn-in set at 10,000,000, sampling every 100,000 states). The combined trees file was annotated in TreeAnnotator v1.8.2 and edited in FigTree v1.4.2. This unrooted maximum credibility tree was presented with height (i.e., ages relative to the youngest sequence), 95% highest posterior density (HPD) intervals, and posterior probabilities placed on the nodes (**Figure 1**).

Pangenome Mining

A pangenome gene presence/absence matrix was generated with an R script based on the annotation spreadsheets extracted using RASTtk (Brettin et al., 2015). Isolates in gene presence/absence matrix were classified in four classes based on geographical origin of isolation (NY, CTX, NTX, and PA) to identify genes that show non-random geospatial distribution. This matrix containing 4,873 genes (Supplementary Table S2) was analyzed using Information gain and ReliefF classification algorithms with default settings in Orange 2.7.8 (Demesar et al., 2013). PCA analysis (“prcomp” method) and Fisher’s exact test (“fisher.test”; e.g., NTX isolates vs. all other) with False Discovery Rate (“p.adjust,” method “hochberg”) were carried out on matrix with excluded constant gene columns (e.g., positive in all isolates or negative in all isolates) in R Studio 0.98, R 3.2.2., package “stats” (R Core Team, 2016). Graphs were plotted using “ggplot” package version 2.1.0 in R (**Figure 2**) (R Core Team, 2016). PlasmidFinder (Carattoli et al., 2014) and PHASTER (Arndt et al., 2016) were used to test for the presence of plasmids and phages in isolates of interest, respectively.

BLAST

Putative virulence genes were extracted from all 86 *S. Cerro* genomes using a standalone BLAST with threshold set at 75% identity and 90% query coverage (Camacho et al., 2009). These virulence genes included D-alanine transporter gene cluster (STM1633 [*dalS*], ST1634 [*dalT*], ST1635 [*dalU*], STM1636 [*dalV*], STM1637), and *sopA* reported in a previous *S. Cerro* study (Rodriguez-Rivera et al., 2014a), as well as 10 genes involved in hydrogen sulfide metabolism (*asrA*, *asrC*, *cysJ*, *cysT*, *phsA*, *phsB*, *phsC*, *ttrA*, *ttrB*, *ttrC*, and STY2774; all from *S. enterica* subsp. *enterica* serotype Typhimurium str. LT2 complete genome deposited on NCBI under gi 16763390).

Hydrogen Sulfide Production

The ability of *S. Cerro* isolates to produce hydrogen sulfide was determined by streaking isolates on agar with thiosulfate

as a source of sulfur [Xylose Lysine Deoxycholate Agar (XLD, BD, East Rutherford, NJ, USA) or Xylose Lysine Tergitol-4 (XLT-4, Northeast Laboratories, Waterville, ME, USA)]. Colonies of isolates able to produce hydrogen sulfide formed a black precipitate after 20–24 h incubation at 35°C due to reaction of hydrogen sulfide with ferric ammonium.

Availability of Data

Trimmed WGS reads were submitted to the SRA under the BioProject ID PRJNA308933. Accession numbers are listed in Supplementary Table S1. Phylogenetic tree (**Figure 1**) file is available on Figshare³ (doi: 10.6084/m9.figshare.4621336). Computational log file and scripts are available in Supplementary Material file log.sh and on GitHub⁴, respectively. Records of NY isolates are available on Food Microbe Tracker⁵. All isolates are available upon request.

RESULTS

The 86 *S. Cerro* isolates from upstate NY (*n* = 54), PA (*n* = 2), CTX (*n* = 27), and NTX (*n* = 3) were whole genome sequenced and analyzed. One of these isolates originated from a human clinical case (NY), two from raw milk (NY), 44 from dairy cattle with and without clinical signs (NY, PA, CTX, NTX), 6 from produce farm environments (NY), and 33 from dairy farm environments (NY, CTX) (Supplementary Table S1). Draft genomes sequenced in this study were assembled with a median of 49 contigs larger than 1 Kb (ranging from 41 to 191), a median N50 of 222,643 (ranging from 39,431 to 352,109 bp), and median average coverage of 144× (ranging from 19× to 406×). The median length of the assembled genomes (made of contigs > 1000 bp) was 4.67 Mbp (ranging from 4.49 to 4.85 Mbp). For all analyzed isolates, assignment to serotype Cerro was confirmed with *in silico* MLST as detailed below. See Supplementary Table S1 for isolate metadata, assembly quality metrics and MLST, and Supplementary Material file log.sh for computational workflow.

Bovine *Salmonella* Cerro Isolates from New York and Texas Belong to a Single MLST Sequence Type and Cluster Geographically Based on Core Genome Sequences

Multilocus sequence typing types of 86 *S. Cerro* isolates were determined with SRST2. A single sequence type (ST 367) was identified for all but three isolates (i.e., FSL R8-7279, FSL R9-1899, and FSL R9-1900) in which we were unable to detect or unambiguously identify one out of seven MLST alleles. In contrast to a number of other *Salmonella* serotypes (e.g., Newport, Kentucky, and Montevideo) (Achtman et al., 2012), a single ST identified in *S. Cerro* confirms the monophyletic nature

³<https://figshare.com/s/aa20239e6e1244026de9>

⁴https://github.com/jasnakovac/salmonella_cerro

⁵<http://www.foodmicrobetracker.com>

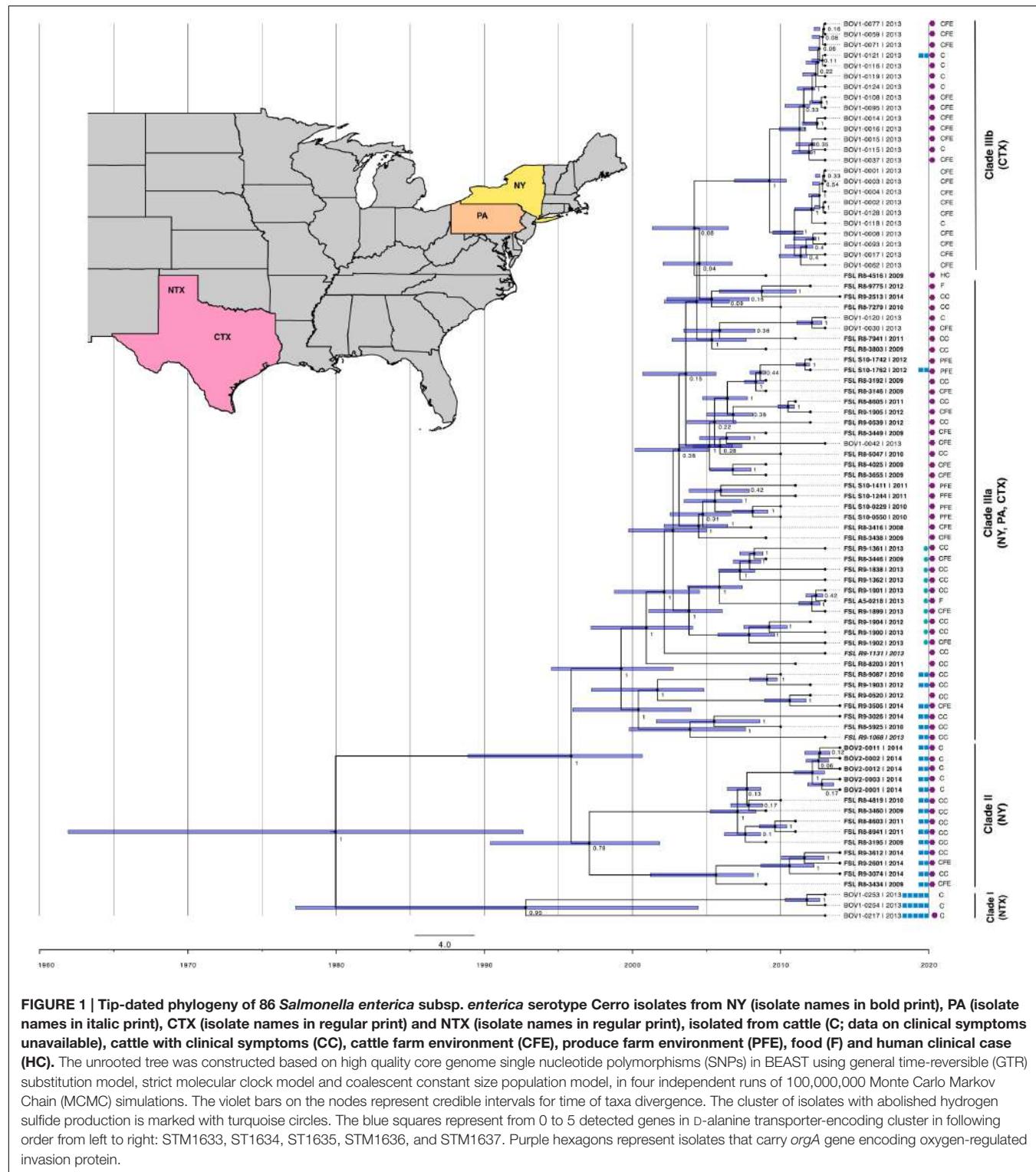


FIGURE 1 | Tip-dated phylogeny of 86 *Salmonella enterica* subsp. *enterica* serotype Cerro isolates from NY (isolate names in bold print), PA (isolate names in italic print), CTX (isolate names in regular print) and NTX (isolate names in regular print), isolated from cattle (C; data on clinical symptoms unavailable), cattle with clinical symptoms (CC), cattle farm environment (CFE), produce farm environment (PFE), food (F) and human clinical case (HC). The unrooted tree was constructed based on high quality core genome single nucleotide polymorphisms (SNPs) in BEAST using general time-reversible (GTR) substitution model, strict molecular clock model and coalescent constant size population model, in four independent runs of 100,000,000 Monte Carlo Markov Chain (MCMC) simulations. The violet bars on the nodes represent credible intervals for time of taxa divergence. The cluster of isolates with abolished hydrogen sulfide production is marked with turquoise circles. The blue squares represent from 0 to 5 detected genes in D-alanine transporter-encoding cluster in following order from left to right: STM1633, ST1634, ST1635, STM1636, and STM1637. Purple hexagons represent isolates that carry *orgA* gene encoding oxygen-regulated invasion protein.

of this serotype, reported in the past (Rodriguez-Rivera et al., 2014a). It also allows for a straightforward ST-based *in silico* serotyping using multilocus or WGSs (Inouye et al., 2014).

We have further resolved phylogenetic relationships among isolates of this monophyletic serotype by identifying 1,434

core genome SNPs using kSNP and building an initial ML phylogeny that guided further analyses. The topology of the ML tree demonstrated clear phylogeographic separation of isolates originating from NY and TX, while the two isolates from PA clustered with NY isolates (Figure 1). Our results suggest that

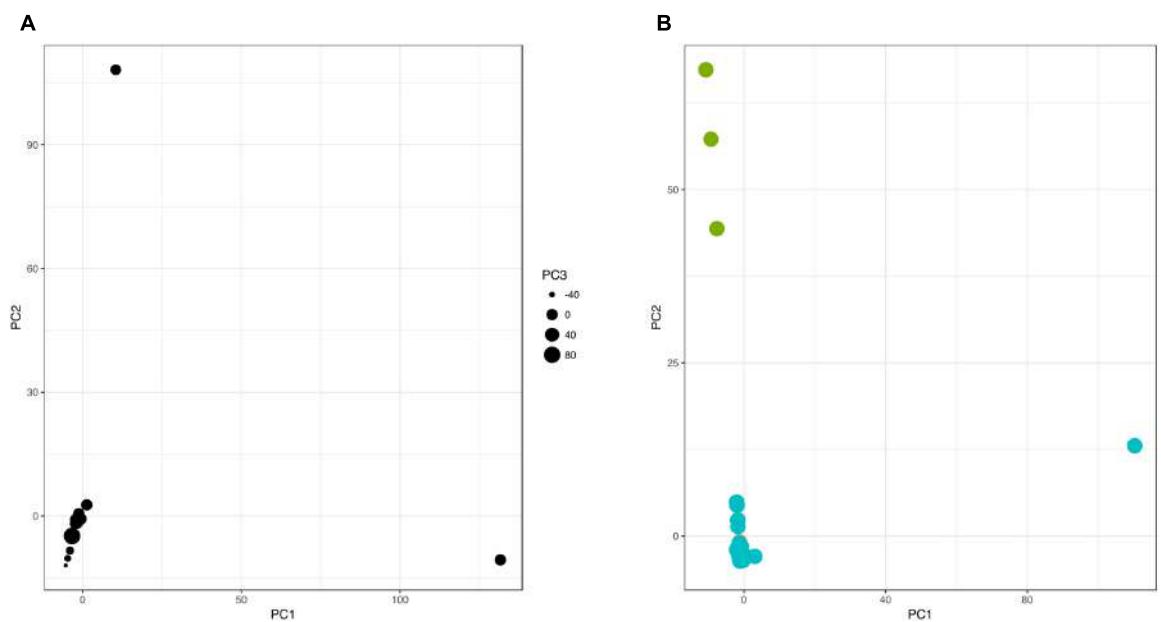


FIGURE 2 | Accessory genome-based PCA clustering of 86 (A) and 84 (B) *Salmonella* Cerro isolates indicates distinct gene patterns of isolates from NTX, compared to isolates from upstate NY, PA, and CTX. PCA analysis was performed in R using prcomp function based on presence/absence data for 1293 genes that were part of an accessory genome of this isolate set. First three principal components were plotted in R using ggplot. The CTX and PA isolates are not visible in the figure, since the dots representing these isolates are covered with the dots representing NY isolates.

CTX isolates appear to have evolved from the northeastern U.S. ancestor (NY or PA) and adapted to dairy cattle in CTX. The CTX genotypes obtained from CTX are only distantly related to NTX.

Bayesian Phylogenetic Reconstruction Suggests Recent Divergence of *S. Cerro* Central Texas Genotype from a Northeastern U.S. Ancestor

To better understand the recent geospatial evolution of *S. Cerro* in the northeast U.S. and Texas, we have carried out detailed variant calling, and used high quality SNPs identified in our set of 86 *S. Cerro* isolates as a base for Bayesian temporal phylogenetic reconstruction. These SNPs were further analyzed in Gubbins to filter out those identified in regions of recombination. Subsequently, 1,319 simple SNPs were used to build a ML phylogeny in RaxML. This phylogenetic tree and isolation years were used to run a linear regression analysis in TempEst to detect the presence of a temporal signal. The correlation coefficient of 0.43, which was determined to be significant using a t-statistic ($P = 1.8^{-5}$), indicated the correlation between isolation date and tip-to-date sequence divergence, which indicates that our dataset is suitable for analysis under the molecular clock assumption. Four different combinations of strict and lognormal relaxed clock models, and coalescent constant and Bayesian skyline models were run in BEAST to reconstruct tip-dated Bayesian phylogeny. Each clock-population model combination was run with a GTR substitution model. Three additional runs were performed with lognormal relaxed clock and coalescent constant model combination, which was identified as the most probable

based on a combination of the tree likelihood, posterior, and ESS (see Table 1). This model combination has also been used in the previous phylogenetic analysis of *S. Cerro* (Rodriguez-Rivera et al., 2014a). The results of all four runs were combined in a single Bayesian tree presented in Figure 1. The 86 isolates included were estimated to evolve with a rate of 7.2×10^{-7} substitutions/site/year (95% HPD $5.2 \times 10^{-7} - 9.3 \times 10^{-7}$).

The 86 analyzed isolates formed three phylogenetic clades (Figure 1). The first clade (clade I; $n = 3$; 2013) comprised NTX isolates, the second clade (clade II; $n = 14$; 2009–2011, 2014) comprised NY isolates, the third clade (clade III) comprised NY, PA, and CTX isolates in a subclade IIIa ($n = 45$; 2008–2014), and only CTX isolates in a subclade IIIb ($n = 24$; 2013). The clade I isolates (NTX) from this study shared a most recent common ancestor (MRCA) with clade II isolates (NY); these clades were estimated to diverge around 1980. The clade II (NY) and clade III (NY, PA, CTX) were estimated to diverge around 1996, which is consistent with the previous study that estimated their MRCA to date back to 1998 (Rodriguez-Rivera et al., 2014a). The NY bovine clade isolates from the aforementioned previous study shared a MRCA with a canine isolate from Washington (isolated in 1989), which evolved from a MRCA shared with feline isolate from Florida (isolated in 1987). The CTX isolates from our study clustered almost exclusively in a subclade IIIb (24/27; 89%) that had evolved from a subclade IIIa comprising NY, PA, and CTX isolates in approximately 2004.

The plausibility of the hypothesis that CTX genotype has evolved from an ancestor originating from the northeast U.S. (NY or PA) was further assessed by examining the available data on proportional prevalence of *S. Cerro* in NY, PA, and TX

TABLE 1 | Mean values and ESSs of key BEAST run statistics computed using GTR substitution model, strict molecular clock, and constant population size models.

Clock model	Population model	Substitution model ^a	Tree	Posterior			Tree model			Constant population size ^c	
				Mean likelihood	ESS ^b	Mean	ESS ^b	Mean	ESS ^b	Mean	ESS ^b
Lognormal relaxed ^d	Constant	GTR	-6.532260E+06	53862	-6.532942E+06	5010	7.21E-07	3272	36	4390	66
	Bayesian skyline	GTR	-6.532256E+06	10179	-6.532202E+06	577	8.73E-07	561	23	746	na
Strict	Constant	GTR	-6.532319E+06	15931	-6.532004E+06	3676	7.28E-07	2616	31	2963	68
	Bayesian skyline	GTR	-6.532318E+06	7905	-6.532308E+06	3050	7.46E-07	2503	29	2958	na

^a GTR, general time-reversible substitution model.^b ESS, effective sample size.^c na, not applicable.^d The parameters shown are a result of four combined BEAST runs initiated from four different random seeds (Supplementary Material file log.sh).

among bovine *Salmonella* isolates. Prevalence of serotype Cerro among *Salmonella* isolated from bovine samples submitted to the Cornell University Animal Health Diagnostic Center started increasing in 2005 from 3% ($n = 20/668$) to 36% ($n = 143/397$) in 2015 (Supplementary Table S2). On the other hand, prevalence of *S. Cerro* isolates among *Salmonella* isolated from bovine samples submitted between 2008 and 2015 to the Veterinary Medical Diagnostic Laboratory in Texas remained relatively stable (Supplementary Table S2).

Accessory Genomes of *S. Cerro* Display a Geospatial Signal

To identify potential genetic traits that may be driving the successful expansion of northeastern U.S. (NY or PA) *S. Cerro* genotype to the south (CTX) of the U.S., we have examined the differences in the accessory genomes of isolates with different geographical origin.

The 86 *S. Cerro* genomes were annotated through RASTtk. Annotation spreadsheets were converted into gene presence/absence matrix (Supplementary Table S3) and analyzed using principle components analysis (PCA) via the prcomp function in RStudio version 0.98.1091, R version 3.3.2 (R Core Team, 2016). The analyzed pangenome comprised 4,873 genes; 3,580 of these were part of a core genome and were therefore removed prior to the PCA analysis of the remaining 1,293 accessory genes. The first 5 and 14 principle components (PCs) captured 50.7 and 71.1% of the cumulative variance, respectively. PC1, PC2, and PC3 were plotted using the ggplot function in ggplot2 2.2.0 (Wickham, 2009). As demonstrated in **Figure 2A**, two isolates, characterized by either a (i) high PC1 and low PC2 value (FSL S10-0550) or (ii) high PC2 and low PC1 value (FSL R8-7941), clustered distinctly compared to the majority of isolates. Isolate FSL S10-0550 was obtained from running water on an upstate NY produce farm in summer 2010, while FSL R8-7941 was isolated from the feces of a bovine case with clinical symptoms in upstate NY in winter 2011. To achieve better separation of the remaining 84 isolates, we have excluded these two isolates and re-run the analysis following the same procedure. This revealed geospatial clustering, separating three NTX isolates from clade I (low PC1, high PC2) from all other isolates (PC1 and PC2 close to 0) (**Figure 2B**).

We observed a gradual loss of genes encoding a putative amino acid ABC transporter, which were detected in all isolates from clades I and II, but only 24% ($n = 11/45$) of isolates from clade IIIa, and 4% ($n = 1/24$) of isolates from clade IIIb (see Supplementary Table S3 and **Figure 1**). In contrast, eight IncI1 plasmid conjugative transfer genes (*pil* and *tra*) were not present in clade I, but were acquired and maintained in all isolates from clades II and III. The same trend was observed for CRISPR repeat with sequence “agtttatccccgtggcgccggaaacac,” which was not detected in clade I, but was gradually enriched in clades II (29%; $n = 4/14$), IIIa (69%; $n = 31/45$), and IIIb (96%; $n = 23/24$). Similarly, we did not detect putative methyltransferase gene in clade I, but did find it in 14% of isolates from clade II, 71% of isolates from clade IIIa, and 96% of isolates from clade IIIb. Another gene, encoding putative cytosine-specific modification

methylase, was found in 30, 7, 73, and 100% of isolates from clades I, II, IIIa, and IIIb, respectively.

To identify groups of genes that are specific for the two distinct isolates from initial PCA analysis (**Figure 2A**), we have identified unique genes contributing to PC1 ($n = 119$) and PC2 ($n = 308$) by comparing the genes contributing to these two PCs before and after excluding these two isolates of interest. The environmental water isolate FSL S10-0550 carried 22 IncF plasmid genes that were not found in genomes of other isolates. Most of these genes were located on contigs smaller than 2 kb. PlasmidFinder identified only one IncFII sequence (pCRY) with 95.28% identity over 593 nt long sequence. This isolate also carried 31 phage genes that were not found in the genomes of other analyzed isolates. The presence of 18 prophages in the genome of FSL S10-0550 was confirmed using PHAST (Zhou et al., 2011). Five of these were intact (i.e., PHAGE_Edward_GF_2_NC_026611, PHAGE_Entero_lato_NC_001422, PHAGE_Entero_P1_NC_005856, PHAGE_Entero_P1_NC_005856, and PHAGE_Phage_Gifsy_1_NC_010392), four were labeled as questionable, and nine as incomplete, based on the completeness score (Zhou et al., 2011). Isolate FSL S10-0550 is phylogenetically very closely related to FSL R8-7941 (**Figure 1**), but has acquired these mobile genetic elements that may be signature for the specific environmental niche. Similarly, isolate FSL R8-7941 had the largest number of IncI1 plasmid genes ($n = 26$), most of which were genes encoding conjugative transfer proteins. Fifteen of these 26 IncI1 genes, as well as 4 IncH1 plasmid genes, were unique to FSL R8-7941. PlasmidFinder identified only one, 142 nt long IncI1 sequence with 100% identity.

Specific Virulence-Associated Genes Were Gradually Lost or Mutated in *S. Cerro*

We found a non-synonymous mutation in a gene encoding SopA in all 86 *S. Cerro* genomes. This mutation resulted in a premature STOP codon on 434th amino acid position in a 782 aa long gene and was found in all 27 *S. Cerro* isolates in a previous study (Rodriguez-Rivera et al., 2014a). Another gene located within *Salmonella* Pathogenicity Island 1 (SPI-1) (*orgA*) was distributed differently among phylogenetic clades. Only one isolate from clade I (NTX) carried this gene, while it was detected in all isolates from clade II (NY) and clade IIIa (NY, PA, CTX) (**Figure 1**). The phylogeny indicates gradual loss of *orgA* in the subclade IIIb (CTX; *orgA* present in 58% isolates; $n = 14/24$).

Isolates Characteristic of Northwest Texas Carry a Full Cluster of D-alanine Transporter Genes, Which Was Gradually Lost in New York and Central Texas Isolates

Next, we investigated the distribution of virulence genes and specific virulence gene variants that have been hypothesized to reduce virulence potential of *S. Cerro* in humans. The D-alanine transporter has been shown to be gradually lost in

S. Cerro from NY in a previous study (Rodriguez-Rivera et al., 2014a). Genes encoding the D-alanine transporter were shown to be required for intracellular survival in murine macrophages (Osborne et al., 2012). The D-alanine transporter is involved in a host-pathogen interaction by limiting D-alanine available to the host neutrophil D-amino acid oxidase, which produces hydrogen peroxide as a side product of D-amino acid metabolism (Tuinema et al., 2014; Takahashi et al., 2015). Bacterial ability to import D-alanine through the D-alanine transporter therefore protects *Salmonella* from oxidative killing mediated by D-amino acid oxidase (Tuinema et al., 2014).

We confirmed the presence of five *S. Typhimurium* LT2 homologs encoding D-alanine transporter genes (STM1633 [*dalS*], ST1634 [*dalT*], ST1635 [*dalU*], and STM1636 [*dalV*]) and STM1637 only in isolates from clade I (**Figure 1**; $n = 3$). Only two out of seven D-alanine transporter genes, (*dalV*) and STM1637, were identified in isolates from clade II. These two genes were detected also in 6 out of 45 (13%) isolates from clade IIIa, and 1 out of 24 (4%) isolates from clade IIIb. The potential impact of D-alanine transporter loss on virulence of *S. Cerro* in cattle remains to be characterized.

Subclade of New York *S. Cerro* Is Characterized by the Inability to Produce Isolation Marker and Virulence Factor Hydrogen Sulfide

Ten isolates (FSL R9-1362, FSL R9-1838, FSL R8-3446, FSL R9-1361, FSL A5-0218, FSL R9-1901, FSL R9-1899, FSL R9-1900, FSL R9-1904, and FSL R9-1902) forming a cluster in a subclade IIIa (**Figure 1**) were not able to produce hydrogen sulfide by reduction of sodium thiosulfate available in XLD and XLT-4 medium. Seven of these isolates possessed 10 genes involved in the hydrogen sulfide metabolic pathway (i.e., *asrC*, *cysJ*, *cyst*, *phsABC*, *ttrABC*, and STY2774). Three isolates, FSL R9-1900, FSL R9-1904, and FSL R9-1902, did not carry *phsA*, which encodes thiosulfate reductase subunit A. The rest of the 10 isolates with impaired ability to produce hydrogen sulfide carried a specific C → T point mutation in *phsA* gene on nucleotide position 1666 of 2277. This mutation resulted in a non-synonymous substitution with a premature stop codon and consequently truncated protein.

DISCUSSION

Phylogenomic analysis of 86 *S. Cerro* isolates from northeast U.S. (NY, PA) and TX indicates that CTX isolates appear to have evolved from the northeastern U.S. ancestor and subsequently adapted to dairy cattle in CTX. The CTX genotypes are only distantly related to endemic NTX genotypes. Our data suggest distinct clustering of NY bovine and environmental *S. Cerro* isolates compared to *S. Cerro* isolates from other sources and geographical regions, which has been shown in a previous study (Rodriguez-Rivera et al., 2014a). The 86 isolates were estimated to evolve with a rate of 7.2×10^{-7} substitutions/site/year (95% HPD 5.2×10^{-7} – 9.3×10^{-7}), which is comparable to the

substitution rate estimated in a previous study of 27 *S. Cerro* isolates (2.4×10^{-7} substitutions/site/year; HPD 1.5×10^{-7} – 3.3×10^{-7}) that were isolated in a broader temporal range (1986–2008 by Rodriguez-Rivera et al., compared to 2008–2014 in the present study) (Rodriguez-Rivera et al., 2014a).

The start of increasing prevalence of *S. Cerro* in NY approximately coincides with the time of a subclinical Cerro outbreak in PA (Van Kessel et al., 2007), as well as with the divergence of the CTX genotype from the NY ancestor in approximately 2004. Overall increasing proportions of this serotype among *Salmonella* have also been identified in samples from clinically ill dairy cattle submitted to the veterinary diagnostic laboratory in central PA between 2005 (14.3%; $N = 33/231$) and 2010 (36.1%; $N = 35/97$) (Tewari et al., 2012). Recently, a cross-sectional study was conducted to estimate the environmental prevalence of *Salmonella* on dairy farms in northwest and CTX (Rodriguez-Rivera et al., 2016). Thirty representative *S. Cerro* isolates from that study were whole genome sequenced in the present study and found to have distinct region-specific genotypes. The NTX genotype (clade I) seems to be endemic in Texas, while the CTX genotype (subclade IIIb) seems to have been introduced to the CTX region in approximately 2004 (Figure 1). Veterinary diagnostic laboratory data from Texas, spanning the years 2008–2015, provide no clear patterns that would suggest an increase in *S. Cerro* prevalence in the CTX region. However, the CTX genotype could have been introduced into a specific region of Texas without generating a detectable increase in prevalence, for example through replacement of another *S. Cerro* genotype or multiple genotypes.

The enrichment of specific CRISPR repeats, IncI1 plasmid conjugative transfer genes and methyltransferase gene in isolates from some phylogenetic clades suggest region-specific adaptation of restriction-modification systems in isolates of serotype Cerro, although restriction-modification systems were recently shown to have limited influence on the overall evolution of *S. enterica* (Roer et al., 2016). The impact of an environment on the bacterial accessory genome was further demonstrated by distinct accessory gene profiles of two isolates originating from environmental water and animal clinical samples. These isolates differed from other isolates by carrying mobile genetic elements, including phage and plasmid genes, which may be linked to specific niches in which they have resided. This demonstrates that the environment can leave specific genomic signatures that are detectable on a fine sub-serotype scale and may be exploited to trace-back the geographic origin of isolates in outbreak investigations.

Furthermore, gradual loss or mutation of virulence genes was observed in isolates from different phylogenetic clades. Examples of such are introduction of premature STOP codon in *sopA* gene in all *S. Cerro* isolates, which has been reported before (Rodriguez-Rivera et al., 2014a), and loss of *orgA* in a subset of isolates from clades I (absent from 2/3 isolates) and IIIb (absent from 10/24 isolates). *sopA* gene is located in one of the major *Salmonella* virulence determinants, SPI-1, which encodes a type III secretion system that allows for the direct delivery of virulence-mediating effector proteins

into the host cell cytoplasm (LaRock et al., 2015). *SopA*, an E6-AP carboxyl terminus (HECT)-like E3 ubiquitin ligase, is an effector protein that functionally mimics at least two mammalian HECT E3 ubiquitin ligases that support the induction of a host immune response, enteritis, and bacterial neutrophil transepithelial migration (Wood et al., 2000; Zhang et al., 2006; Kamanova et al., 2016). E3 ubiquitin ligase determines the specificity of proteins destined to undergo ubiquitination, which is essential for a number of cellular functions that involve protein degradation (Zhang et al., 2006). *SopA* was shown to be involved in induction of diarrhea in calves infected with *S. Typhimurium* via the oral route (Zhang et al., 2002), but it is not known yet whether and how the truncation of *SopA* influences the virulence of *S. Cerro* in cattle, as the isolates studied here were obtained from both cattle with and without clinical evidence of disease, as well as the environment. In contrast to *sopA*, gene encoding *OrgA* was absent from a subset of Texas isolates (2/3 isolates in clade I and 10/24 isolates in clade IIIb), but was found in all isolates from NY and PA (clades II and IIIa). *OrgA* was initially shown to play an important role in invasion of murine cells under low-oxygen conditions when *Salmonella* is administered through an oral route (Jones and Falkow, 1994). Importantly, *OrgA* was shown to be essential for type 3 secretion system assembly, as *S. Typhimurium* mutants with inactivated *orgA* do not form a needle substructure, which is necessary for formation of a functional secretion system, and invasion in epithelial cells (Sukhan et al., 2001). The loss of *orgA* in a subset of Texas isolates may suggest decreased virulence in cattle for this genotype. However, we were not able to confirm this without specific data on clinical signs or extent of clinical illness among cattle sampled in Texas, as not enough metadata were available.

We observed gradual loss of a gene encoding a D-alanine amino acid ABC transporter, which has been reported previously (Rodriguez-Rivera et al., 2014a). D-alanine ABC transporter is one of *Salmonella* virulence factors, and its gradual loss in NY, PA, and CTX isolates suggests temporal adaptation of *S. Cerro* to a bovine host, likely allowing for its successful spread among cattle.

Another phenomenon observed among a subset of analyzed *S. Cerro* isolates was loss of ability to produce hydrogen sulfide, a virulence factor and microbiological isolation marker, due to mutation causing a premature STOP codon in a *pshA* gene encoding thiosulfate reductase subunit A. This specific point mutation has been associated with impaired ability to produce hydrogen sulfide in another study that investigated Japanese *S. Typhimurium* and *S. Infantis* isolates from poultry meat (Sakano et al., 2013). The C → T mutation was also found in *S. Aberdeen* food isolates from China ($n = 7/160$; 4.4%), but on a 208th position (Wu et al., 2016). The positions may differ, however, due to a different reference sequence length used when comparing PCR products or full-length gene extracted from WGS. Another Chinese study reported non-hydrogen-sulfide-producing *S. enterica* subsp. *enterica* found in chicken ($n = 20/29$; 69%; predominantly *S. Derby* and *S. Heidelberg*) and pork meat samples ($n = 13/53$; 25%) (Lin et al., 2014). Emergence of *Salmonella* isolates with an impaired ability to produce

hydrogen sulfide is concerning, as this phenotype increases the risk for false negative detection of *Salmonella* using traditional microbiological methods, which rely on characteristic black color of hydrogen sulfide precipitate on selective differential media. Furthermore, the ability of *Salmonella* to reduce tetrathionate to hydrogen sulfite in a host gastrointestinal tract provides it with a competitive growth advantage over microbiota that cannot exploit tetrathionate as an alternative electron acceptor (Winter et al., 2010). The influence of H₂S-negative phenotype may therefore decrease *Salmonella* virulence in a host. The key role of tetrathionate reductase subunit A-encoding gene (*ttrA*) in providing growth advantage in a host was demonstrated in the Winter et al. (2010) study, while the direct involvement of *phsA* remains to be confirmed. A recent study found two phylogenetically distinct clades of *S. Senftenberg* outbreak isolates from China, SC1 and SC2. Isolates belonging to the SC1 clade carried a different variant of SPI-1 and were less invasive and not able to produce hydrogen sulfide; however, the relative contribution of these traits to pathogenicity is not yet understood (Abd El Ghany et al., 2016). The hydrogen sulfide negative isolates from our study were isolated both from cattle with and without clinical signs, as well as the dairy farm environment, precluding us from drawing conclusions about association of this phenotype with virulence in cattle.

CONCLUSION AND IMPLICATIONS

Core genome-based temporal reconstruction of phylogenetic relationships among 86 *S. Cerro* isolates from NY, PA, and Texas suggests recent transmission and divergence of CTX genotype from a northeastern U.S. ancestor. Several genomic markers associated with geographic origin suggest the ability of *Salmonella* to rapidly diverge and adapt to specific niches, and to modify virulence-related characteristics, such as the ability to fully utilize tetrathionate as an alternative electron acceptor, which is commonly used to detect *Salmonella*. Increased proportional prevalence of this serotype among *Salmonella* isolates from clinical dairy cattle samples in a number of

U.S. states demonstrates the need for development of control strategies to effectively mitigate the transmission of this serotype. Furthermore, a cluster of isolates that are not able to produce H₂S suggests the emergence of *Salmonella* strains with this phenotype, which is challenging to detect using traditional microbiological methods for detection of *Salmonella*. Overall, our data show that clinical outcome data and genetic data for *S. Cerro* isolates, such as truncations in virulence genes leading to novel phenotypes, should be correlated to allow for accurate risk assessment.

AUTHOR CONTRIBUTIONS

JK and LC performed computational and statistical data analyses; LR-R performed experimental analyses. AT contributed the data. JK, KC, and MW conceived the study. JK and MW co-wrote the manuscript.

FUNDING

This material is based upon work that is partially supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, Hatch under project NYC-143436, and the Texas A&M Genomics Seed Grant Program.

ACKNOWLEDGMENTS

The authors thank Steven Warchocki for genomic DNA preparation, Jeffrey I. Tokman for evaluation of hydrogen sulfide production, and Daniel L. Weller for assistance with map design.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.00737/full#supplementary-material>

REFERENCES

- Abd El Ghany, M., Shi, X., Li, Y., Ansari, H. R., Hill-Cawthorne, G. A., Ho, Y. S., et al. (2016). Genomic and phenotypic analyses reveal the emergence of an atypical *Salmonella enterica* serovar senftenberg variant in China. *J. Clin. Microbiol.* 54, 2014–2022. doi: 10.1128/JCM.00052-16
- Achtman, M., Wain, J., Weill, F. X., Nair, S., Zhou, Z., Sangal, V., et al. (2012). Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog.* 8:e1002776. doi: 10.1371/journal.ppat.1002776
- Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi: 10.1093/nar/gkw387
- Babraham Bioinformatics (2014). *FastQC v. 0.11.2*. Available at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., et al. (2015). RASTk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* 5:8365. doi: 10.1038/srep08365
- Bushnell, B. (2015). *BBMap v. 35.49*. Available at: <https://sourceforge.net/projects/bbmap/>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421
- Carattoli, A., Zankari, E., Garcia-Fernandez, A., Voldby Larsen, M., Lund, O., Villa, L., et al. (2014). *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* 58, 3895–3903. doi: 10.1128/AAC.02412-14
- Chapagain, P. P., van Kessel, J. S., Karns, J. S., Wolfgang, D. R., Hovingh, E., Nelen, K. A., et al. (2008). A mathematical model of the dynamics of *Salmonella* Cerro

- infection in a US dairy herd. *Epidemiol. Infect.* 136, 263–272. doi: 10.1017/S0950268807008400
- Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., et al. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* 43, e15. doi: 10.1093/nar/gku1196
- Cummings, K. J., Warnick, L. D., Elton, M., Grohn, Y. T., McDonough, P. L., and Siler, J. D. (2010a). The effect of clinical outbreaks of salmonellosis on the prevalence of fecal *Salmonella* shedding among dairy cattle in New York. *Foodborne Pathog. Dis.* 7, 815–823. doi: 10.1089/fpd.2009.0481
- Cummings, K. J., Warnick, L. D., Elton, M., Rodriguez-Rivera, L. D., Siler, J. D., Wright, E. M., et al. (2010b). *Salmonella enterica* serotype Cerro among dairy cattle in New York: an emerging pathogen? *Foodborne Pathog. Dis.* 7, 659–665. doi: 10.1089/fpd.2009.0462
- Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., et al. (2013). Orange: data mining toolbox in python. *J. Mach. Learn. Res.* 14, 2349–2353.
- Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUTi and the BEAST 1.7. *Mol. Biol. Evol.* 29, 1969–1973. doi: 10.1093/molbev/mss075
- Gardner, S. N., and Hall, B. G. (2013). When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PLoS ONE* 8:e81760. doi: 10.1371/journal.pone.0081760
- Hong, S., Rovira, A., Davies, P., Ahlstrom, C., Muellner, P., Rendahl, A., et al. (2016). Serotypes and antimicrobial resistance in *Salmonella enterica* recovered from clinical samples from cattle and Swine in Minnesota, 2006 to 2015. *PLoS ONE* 11:e0168016. doi: 10.1371/journal.pone.0168016
- Inouye, M., Dashnow, H., Raven, L. A., Schultz, M. B., Pope, B. J., Tomita, T., et al. (2014). SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* 6:90. doi: 10.1186/s13073-014-0090-6
- Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* 44, 226–232. doi: 10.1038/ng.1028
- Jones, B. D., and Falkow, S. (1994). Identification and characterization of a *Salmonella* Typhimurium oxygen-regulated gene required for bacterial internalization. *Infect. Immun.* 62, 3745–3752.
- Jones, T. F., Ingram, L. A., Cieslak, P. R., Vugia, D. J., Tobin-D'Angelo, M., Hurd, S., et al. (2008). Salmonellosis outcomes differ substantially by serotype. *J. Infect. Dis.* 198, 109–114. doi: 10.1086/588823
- Kamanova, J., Sun, H., Lara-Tejero, M., and Galan, J. E. (2016). The *Salmonella* effector protein SopA modulates innate immune responses by targeting TRIM E3 ligase family members. *PLoS Pathog.* 12:e1005552. doi: 10.1371/journal.ppat.1005552
- LaRock, D. L., Chaudhary, A., and Miller, S. I. (2015). *Salmonellae* interactions with host processes. *Nat. Rev. Microbiol.* 13, 191–205. doi: 10.1038/nrmicro3420
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Lin, D., Yan, M., Lin, S., and Chen, S. (2014). Increasing prevalence of hydrogen sulfide negative *Salmonella* in retail meats. *Food Microbiol.* 43, 1–4. doi: 10.1016/j.fm.2014.04.010
- Osborne, S. E., Tuinema, B. R., Mok, M. C., Lau, P. S., Bui, N. K., Tomljenovic-Berube, A. M., et al. (2012). Characterization of DalS, an ATP-binding cassette transporter for D-alanine, and its role in pathogenesis in *Salmonella enterica*. *J. Biol. Chem.* 287, 15242–15250. doi: 10.1074/jbc.M112.348227
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>
- Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* 2:vew007. doi: 10.1093/ve/vew007
- Rodriguez-Rivera, L. D., Cummings, K. J., Loneragan, G. H., Rankin, S. C., Hanson, D. L., Leone, W. M., et al. (2016). *Salmonella* prevalence and antimicrobial susceptibility among dairy farm environmental samples collected in Texas. *Foodborne Pathog. Dis.* 13, 205–211. doi: 10.1089/fpd.2015.2037
- Rodriguez-Rivera, L. D., Moreno Switt, A. I., Degoricija, L., Fang, R., Cummings, C. A., Furtado, M. R., et al. (2014a). Genomic characterization of *Salmonella* Cerro ST367, an emerging *Salmonella* subtype in cattle in the United States. *BMC Genomics* 15:427. doi: 10.1186/1471-2164-15-427
- Rodriguez-Rivera, L. D., Wright, E. M., Siler, J. D., Elton, M., Cummings, K. J., Warnick, L. D., et al. (2014b). Subtype analysis of *Salmonella* isolated from subclinically infected dairy cattle and dairy farm environments reveals the presence of both human- and bovine-associated subtypes. *Vet. Microbiol.* 170, 307–316. doi: 10.1016/j.vetmic.2014.02.013
- Roer, L., Hendriksen, R. S., Leekitcharoenphon, P., Lukjancenko, O., Kaas, R. S., Hasman, H., et al. (2016). Is the evolution of *Salmonella enterica* subsp. *enterica* Linked to Restriction-Modification Systems?. *mSystems* 1:e00009–e16. doi: 10.1128/mSystems.00009-16
- Sakano, C., Kuroda, M., Sekizuka, T., Ishioka, T., Morita, Y., Ryo, A., et al. (2013). Genetic analysis of non-hydrogen sulfide-producing *Salmonella enterica* serovar Typhimurium and S. enterica serovar Infantis isolates in Japan. *J. Clin. Microbiol.* 51, 328–330. doi: 10.1128/JCM.02225-12
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033
- Sukhan, A., Kubori, T., Wilson, J., and Galan, J. E. (2001). Genetic analysis of assembly of the *Salmonella enterica* serovar Typhimurium type III secretion-associated needle complex. *J. Bacteriol.* 183, 1159–1167. doi: 10.1128/JB.183.4.1159-1167.2001
- Takahashi, S., Abe, K., and Kera, Y. (2015). Bacterial d-amino acid oxidases: recent findings and future perspectives. *Bioengineered* 6, 237–241. doi: 10.1080/21655979.2015.1052917
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197
- Tewari, D., Sandt, C. H., Miller, D. M., Jayarao, B. M., and M'Ikanatha, N. M. (2012). Prevalence of *Salmonella* Cerro in laboratory-based submissions of cattle and comparison with human infections in Pennsylvania, 2005–2010. *Foodborne Pathog. Dis.* 9, 928–933. doi: 10.1089/fpd.2012.1142
- Tuinema, B. R., Reid-Yu, S. A., and Coombes, B. K. (2014). *Salmonella* evades D-amino acid oxidase to promote infection in neutrophils. *MBio* 5:e01886. doi: 10.1128/mBio.01886-14
- United States Department of Agriculture [USDA] (2011). *Salmonella, Listeria, and Campylobacter on U. S. Dairy Operations*, 1996–2007. Fort Collins, CO: Centers for Epidemiology and Animal Health.
- Valenzuela, J. R., Sethi, A. K., Aulik, N. A., and Poulsen, K. P. (2017). Antimicrobial resistance patterns of bovine *Salmonella enterica* isolates submitted to the Wisconsin Veterinary Diagnostic Laboratory: 2006–2015. *J. Dairy Sci.* 100, 1319–1330. doi: 10.3168/jds.2016-11419
- Van Kessel, J. A., Karns, J. S., Wolfgang, D. R., and Hovingh, E. (2013). Regional distribution of two dairy-associated *Salmonella enterica* serotypes. *Foodborne Pathog. Dis.* 10, 448–452. doi: 10.1089/fpd.2012.1380
- Van Kessel, J. S., Karns, J. S., Wolfgang, D. R., Hovingh, E., and Schukken, Y. H. (2007). Longitudinal study of a clonal, subclinical outbreak of *Salmonella enterica* subsp. *enterica* serovar Cerro in a U.S. dairy herd. *Foodborne Pathog. Dis.* 4, 449–461. doi: 10.1089/fpd.2007.0033
- Vangay, P., Fugett, E. B., Sun, Q., and Wiedmann, M. (2013). Food microbe tracker: a web-based tool for storage and comparison of food-associated microbes. *J. Food Prot.* 76, 283–294. doi: 10.4315/0362-028X.JFP-12-276
- Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer. doi: 10.1007/978-0-387-98141-3
- Winter, S. E., Thiennimitr, P., Winter, M. G., Butler, B. P., Huseby, D. L., Crawford, R. W., et al. (2010). Gut inflammation provides a respiratory electron acceptor for *Salmonella*. *Nature* 467, 426–429. doi: 10.1038/nature09415
- Wood, M. W., Jones, M. A., Watson, P. R., Siber, A. M., McCormick, B. A., Hedges, S., et al. (2000). The secreted effector protein of *Salmonella* Dublin, SopA, is translocated into eukaryotic cells and influences the induction of enteritis. *Cell Microbiol.* 2, 293–303. doi: 10.1046/j.1462-5822.2000.00054.x
- Wu, F., Xu, X., Xie, J., Yi, S., Wang, J., Yang, X., et al. (2016). Molecular characterization of *Salmonella enterica* serovar aberdeen negative for H2S production in China. *PLoS ONE* 11:e0161352. doi: 10.1371/journal.pone.0161352

- Zhang, S., Santos, R. L., Tsolis, R. M., Stender, S., Hardt, W. D., Baumler, A. J., et al. (2002). The *Salmonella enterica* serotype Typhimurium effector proteins SipA, SopA, SopB, SopD, and SopE2 act in concert to induce diarrhea in calves. *Infect. Immun.* 70, 3843–3855. doi: 10.1128/IAI.70.7.3843-3855.2002
- Zhang, Y., Higashide, W. M., McCormick, B. A., Chen, J., and Zhou, D. (2006). The inflammation-associated *Salmonella* SopA is a HECT-like E3 ubiquitin ligase. *Mol. Microbiol.* 62, 786–793. doi: 10.1111/j.1365-2958.2006.05407.x
- Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Res.* 39, W347–W352. doi: 10.1093/nar/gkr485

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Kovac, Cummings, Rodriguez-Rivera, Carroll, Thachil and Wiedmann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Metagenomics: The Next Culture-Independent Game Changer

Jessica D. Forbes^{1,2†}, Natalie C. Knox^{1†}, Jennifer Ronholm^{3,4}, Franco Pagotto^{5,6} and Aleisha Reimer^{1*}

¹ National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada, ² Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, MB, Canada, ³ Department of Food Science and Agricultural Chemistry, Faculty of Agricultural and Environmental Sciences, McGill University, Montreal, QC, Canada, ⁴ Department of Animal Science, Faculty of Agricultural and Environmental Sciences, McGill University, Montreal, QC, Canada, ⁵ Bureau of Microbial Hazards, Food Directorate, Health Canada, Ottawa, ON, Canada, ⁶ Listeriosis Reference Centre, Bureau of Microbial Hazards, Food Directorate, Health Canada, Ottawa, ON, Canada

OPEN ACCESS

Edited by:

David Rodriguez-Lazaro,
University of Burgos, Spain

Reviewed by:

Dario De Medici,
Istituto Superiore di Sanità, Italy
Beatrix Stessl,
Veterinärmedizinische Universität
Wien, Austria

*Correspondence:

Aleisha Reimer
aleisha.reimer@phac-aspc.gc.ca

[†]These authors have contributed
equally to this work.

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 20 February 2017

Accepted: 29 May 2017

Published: 04 July 2017

Citation:

Forbes JD, Knox NC, Ronholm J, Pagotto F and Reimer A (2017)
Metagenomics: The Next Culture-Independent Game Changer.

Front. Microbiol. 8:1069.
doi: 10.3389/fmicb.2017.01069

A trend towards the abandonment of obtaining pure culture isolates in frontline laboratories is at a crossroads with the ability of public health agencies to perform their basic mandate of foodborne disease surveillance and response. The implementation of culture-independent diagnostic tests (CIDTs) including nucleic acid and antigen-based assays for acute gastroenteritis is leaving public health agencies without laboratory evidence to link clinical cases to each other and to food or environmental substances. This limits the efficacy of public health epidemiology and surveillance as well as outbreak detection and investigation. Foodborne outbreaks have the potential to remain undetected or have insufficient evidence to support source attribution and may inadvertently increase the incidence of foodborne diseases. Next-generation sequencing of pure culture isolates in clinical microbiology laboratories has the potential to revolutionize the fields of food safety and public health. Metagenomics and other ‘omics’ disciplines could provide the solution to a cultureless future in clinical microbiology, food safety and public health. Data mining of information obtained from metagenomics assays can be particularly useful for the identification of clinical causative agents or foodborne contamination, detection of AMR and/or virulence factors, in addition to providing high-resolution subtyping data. Thus, metagenomics assays may provide a universal test for clinical diagnostics, foodborne pathogen detection, subtyping and investigation. This information has the potential to reform the field of enteric disease diagnostics and surveillance and also infectious diseases as a whole. The aim of this review will be to present the current state of CIDTs in diagnostic and public health laboratories as they relate to foodborne illness and food safety. Moreover, we will also discuss the diagnostic and subtyping utility and concomitant bias limitations of metagenomics and comparable detection techniques in clinical microbiology, food and public health laboratories. Early advances in the discipline of metagenomics, however, have indicated noteworthy challenges. Through forthcoming improvements in sequencing technology and analytical pipelines among others, we anticipate that within the next decade, detection and characterization of pathogens via metagenomics-based workflows will be implemented in routine usage in diagnostic and public health laboratories.

Keywords: metagenomics, targeted-amplicon, food safety, public health, culture-independent diagnostic test, next-generation sequencing, antimicrobial resistance, molecular epidemiology

INTRODUCTION

The incidence and impact of foodborne illness constitutes a significant global issue to public health. Foodborne illness affects one in eight Canadians annually, resulting in an estimated 4 million infections, 11,600 hospitalizations and 238 deaths (Thomas et al., 2015); the leading causes of known foodborne infections include norovirus (65%), *Clostridium perfringens* (11%), *Campylobacter* spp. (8%), and non-typhoidal *Salmonella* spp. (5%). According to Health Canada, approximately 2.4 million cases or 60% of foodborne illness are attributed to unknown causes versus only 1.6 million cases or 40% causatively linked to 30 recognized foodborne microbes which include bacteria, viruses and parasites (Thomas et al., 2015). The detection of foodborne enteric pathogens and hence diagnosis of foodborne disease, historically (and still considered the gold standard) have been conducted via culture-dependent techniques, that is, the physical isolation of a bacterial pathogen. Though recognized for some time, the utilization of CIDTs has been increasing throughout the last decade, effectively transforming clinical and food microbiology laboratories (Janda and Abbott, 2014).

In clinical diagnostic laboratories, pathogen detection is increasingly contingent upon the analytic application of CIDTs, which include nucleic acid (e.g., PCR) and antigen-based tests (e.g., ELISA) among others. The extensive adaptation of CIDTs to clinical and food settings is largely reliant on inherent advantages over traditional culture-dependent diagnostic tests; use of CIDTs offers a considerably faster TAT which is crucial for (i) clinical decision-making and decreasing the unnecessary use of broad-spectrum or ineffective antimicrobials, (ii) early outbreak detection and control, and (iii) food industry release or recall of products. Moreover, most conventional CIDTs require less technical expertise and in the long-term may offer a more cost effective alternative.

The clinical use of CIDTs presents the potential to improve disease detection. First, CIDTs are reportedly more sensitive and specific than culture (Hanson and Couturier, 2016; Steyer et al., 2016). Second, reduced complexity of CIDTs allows for rapid testing thereby allowing for a higher throughput of biological specimens to be tested. Third, CIDTs can identify non-culturable or fastidious microbes such as *Campylobacter* spp. (Fitzgerald et al., 2016) or noroviruses (Jones et al., 2014). Lastly, CIDTs enhance the ability to identify polymicrobial or complex infections. Not only are CIDTs attractively useful in

the clinical microbiology laboratory but also in recent years their utilization has become widespread for routine and rapid detection of common pathogens in other sectors including public health and food safety laboratories in addition to *in situ* testing of food processing establishments and agricultural sites. Traditional diagnostics, however, rely upon tests that are tailored to the etiological agent associated with a particular syndrome.

Next-generation sequencing technology has effectively transformed infectious disease research throughout the last decade. High-throughput laboratory techniques are bypassing onerous testing via complement or replacement of conventional microbiological, molecular and serological tests for identifying, typing and characterizing pathogens. WGS of cultured isolates has been extensively employed for pathogen characterization, outbreak detection, phylogenomics and microbial genome wide association studies and thus has progressed from the proof-of-principle phase to implementation in routine foodborne surveillance and outbreak response (Jackson et al., 2016).

Current public health infectious disease surveillance methodologies are generally reliant upon the frontline laboratory to refer pure culture isolates to their local or provincial public health laboratory. As we progressively enter a culture-independent era for infectious disease diagnostics, public health laboratories will inevitably receive fewer isolates. Hence, performing appropriate subtyping and AST assays to identify and track foodborne outbreaks will be difficult. Inadequate surveillance measures have the potential to negatively affect food safety via the inability to identify outbreaks and perform source attribution studies of contaminated foods. We expect this will result in an increase of contaminated foods lingering on the market thereby causing more cases of undetected foodborne outbreaks (Carleton and Gerner-Smidt, 2016). **Figure 1** shows a schematic representation of the effect of CIDTs on the food industry.

Employment of molecular methods that are highly informative will improve the all-encompassing issues associated between primary diagnostic and public health laboratories. Metagenomics for example offers the advantage of a less biased pathogen detection methodology through direct sequencing of the specimen's extracted DNA. This approach has the potential to capture a thorough representation of the microbial community (with some limitations; discussed below) thus eliminating the requirement for pure culture. Metagenomics and similar techniques traditionally have been applied to interrogate microbiomes of a particular ecological niche through sequencing of all nucleic acids recovered from a sample (Eckburg et al., 2005; Huttenhower et al., 2012; Mason et al., 2014). Moreover, a number of clinically relevant applications stand to benefit from such data – rapid identification of the etiological agent (known or novel) and gene content including virulence and AMR, or inferring functional pathways to elucidate multifaceted illnesses.

Laboratory methods that detect and identify pathogens serve two critical functions: clinical decision-making (individual level) and public health decision-making (population level). While traditional culture-based methods meet both of these needs simultaneously, CIDTs are exclusively designed to improve

Abbreviations: AMR, antimicrobial resistance; AST, antimicrobial susceptibility testing; CIDT, culture independent diagnostic test; CSF, cerebrospinal fluid; DNA, deoxyribonucleic acid; ELISA, enzyme linked immunosorbent assay; FDA, food and drug administration; FISH, fluorescent *in situ* hybridization; GMI, global microbial identifier; ITS, internal transcribed spacer; LAMP, loop mediated isothermal amplification; LFA, lateral flow assay; LLS, listeriolysin S; MALDI-TOF, matrix assisted laser desorption/ionization time of flight; MDR, multidrug-resistant; MLST, multilocus sequencing typing; MLVA, multilocus variable number tandem repeat analysis; NAAT, nucleic acid amplification test; NASBA, nucleic acid sequence based amplification; NGS, next-generation sequencing; PCR, polymerase chain reaction; PFGE, pulsed field gel electrophoresis; rRNA, ribosomal ribonucleic acid; SNV, single nucleotide variant; STEC, shiga-toxin producing *Escherichia coli*; SURPI, sequence-based ultra-rapid pathogen identification; TAT, turn-around-time; WGS, whole genome sequencing.

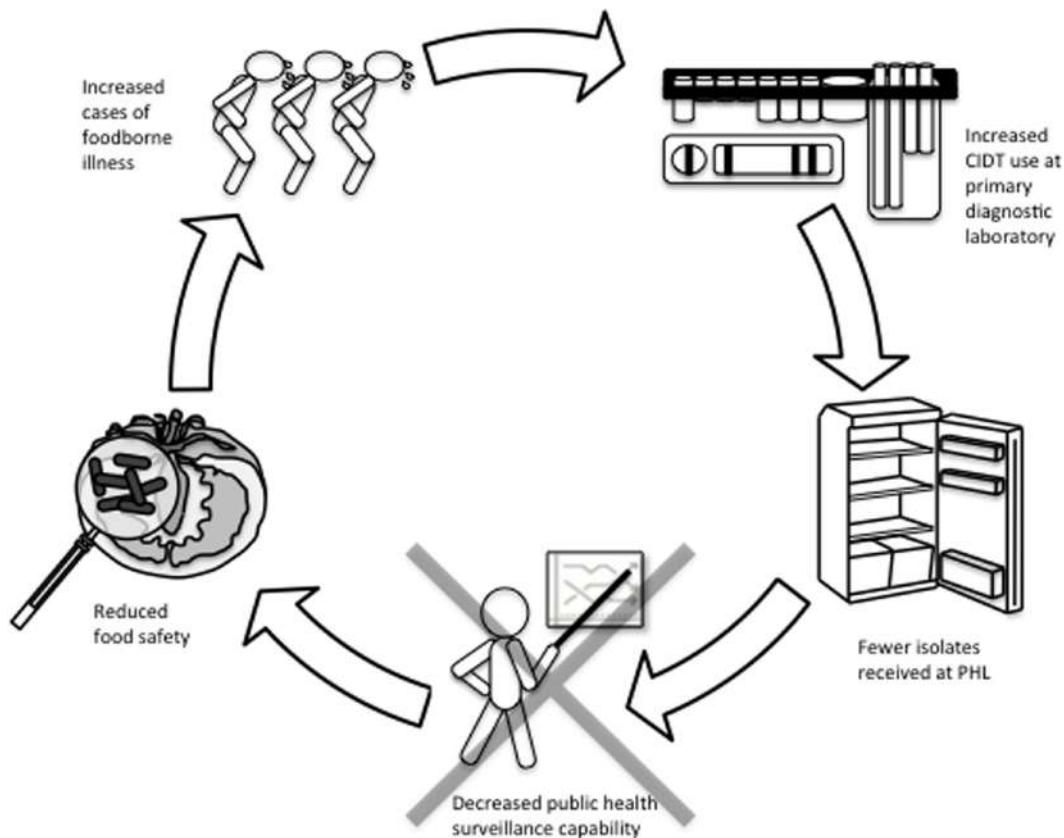


FIGURE 1 | Cycle of how CIDT upsurge in clinical laboratories may have compound effects on the food industry. Increased usage of culture-independent syndromic panels to diagnose cases of acute gastroenteritis is expected to result in fewer microbial cultures referred to public health laboratories. Fewer isolates at public health laboratories will limit surveillance abilities with negative impacts to (i) monitoring trends in AMR, (ii) detection and response to food safety incidents and outbreaks, and (iii) source attribution. Consequently, the food supply in Canada and potentially other countries, may be exposed to an increase in foodborne illness. In turn, the repercussions of this cycle in the absence of reflex cultures could have serious implications at all levels of foodborne disease management.

clinical decision making only. This leaves critical public health activities at great risk. A summary of techniques used in primary diagnostic labs is illustrated in **Figure 2**. Herein, we will review the current state of CIDTs as they relate to foodborne illness and food safety. We will furthermore discuss the diagnostic utility of metagenomics and comparable detection techniques in clinical microbiology, food and public health laboratories.

CURRENT STATE OF CONVENTIONAL CULTURE-INDEPENDENT DIAGNOSTIC TESTS IN CLINICAL AND FOOD MICROBIOLOGY LABORATORIES

Traditional culture-based methods for detecting enteric pathogens in biological specimens, food processing environments or foods is dependent on the growth of viable and culturable microbes; biochemical tests and subtyping are further needed to confirm microbe identification and to establish genetic linkage, respectively. Though sensitive, relatively easy

and inexpensive, culture-based methods are laborious and ineffective for non-cultivable or fastidious microbes. The diagnostic potential of CIDTs to circumvent the demanding task of culturing and subtyping pathogens is a widely attractive alternative.

Conventional Culture-Independent Diagnostic Tests

Though globally, data is limited regarding the current use of CIDTs in clinical and food laboratories, many clinical laboratories in the US are reportedly in the process of converting to the use of CIDTs for the identification of enteric microbes (Shea et al., 2017). According to the CDC, significant increases in the amount of enteric infections diagnosed in the US entirely by CIDTs were reported in 2015 compared to year's prior (2012–2014). In particular, increases in positive CIDTs were reported for 4 enteric pathogens: *Campylobacter* (92%), *Shigella* (284%), *Salmonella* (247%), and STEC (120%; Huang et al., 2016). We expect the increased use of CIDTs is less dramatic in Canada, with suspected regional differences, due to health-care system differences, though there is a paucity

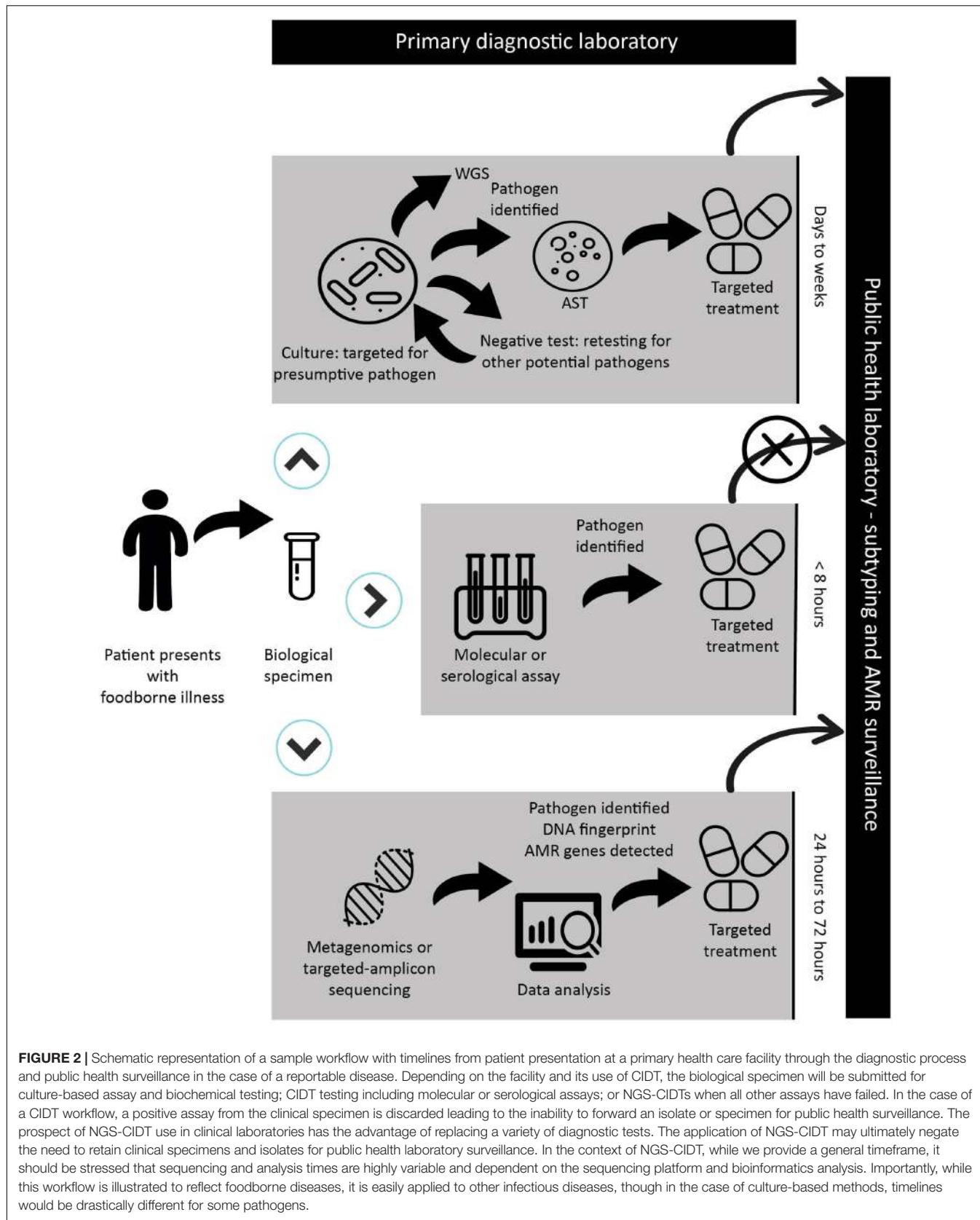


FIGURE 2 | Schematic representation of a sample workflow with timelines from patient presentation at a primary health care facility through the diagnostic process and public health surveillance in the case of a reportable disease. Depending on the facility and its use of CIDT, the biological specimen will be submitted for culture-based assay and biochemical testing; CIDT testing including molecular or serological assays; or NGS-CIDTs when all other assays have failed. In the case of a CIDT workflow, a positive assay from the clinical specimen is discarded leading to the inability to forward an isolate or specimen for public health surveillance. The prospect of NGS-CIDT use in clinical laboratories has the advantage of replacing a variety of diagnostic tests. The application of NGS-CIDT may ultimately negate the need to retain clinical specimens and isolates for public health laboratory surveillance. In the context of NGS-CIDT, while we provide a general timeframe, it should be stressed that sequencing and analysis times are highly variable and dependent on the sequencing platform and bioinformatics analysis. Importantly, while this workflow is illustrated to reflect foodborne diseases, it is easily applied to other infectious diseases, though in the case of culture-based methods, timelines would be drastically different for some pathogens.

of published data to support this theory. For example, the Canadian health care system is not driven by private insurance, as is the case in the US, but rather, is publicly funded. To establish trends of CIDT uptake in Canada, surveys will be required to evaluate their usage in primary diagnostic laboratories. We anticipate that the percentage of infections diagnosed with CIDTs will continue to rise in clinical and food microbiology laboratories in developed countries in the coming years.

This past decade has seen a drastic expansion in the diagnostic application of available and validated CIDTs (Doggett et al., 2016; Shea et al., 2017). There is a wide degree of variation amongst implementation of CIDTs in diagnostic laboratories ranging from singleplex PCR assays to complete laboratory automation; thus, microbiologists are able to yield clinically actionable results of superior quality ultimately benefiting patient health and outcome. Conventional CIDTs can be assigned to one of two methodologies: nucleic acid-based methods and antigen-based tests. The use of nucleic acid-based assays to identify enteric pathogens presents with several advantages over culture-dependent tests in clinical and food settings. Their use allows for high levels of sensitivity and specificity and offers an added benefit in their ability to detect toxin-producing genes or other important biomarkers. Numerous nucleic acid assays are in routine use including PCR (singleplex, multiplex, quantitative, quantitative reverse-transcription, real-time), amplification such as LAMP and NASBA, DNA microarray, microfluidic chip and MALDI-TOF mass spectrometry. Alternatively, antigen-based tests for enteric pathogen detection include ELISA and LFA; reviewed in Zhao et al. (2014). A detailed description of conventional CIDTs is outside the scope of this review and we refer readers to an exhaustive review of CIDTs (Doggett et al., 2016).

A growing number of US FDA – approved syndromic panels for multiple pathogen detection in addition to laboratory-developed tests have facilitated the upsurge of CIDT application for acute gastroenteritis in clinical laboratories. These panels employ PCR to detect unique DNA sequences to enable identification of enteric pathogens. Several commercially available gastrointestinal multiplex panels are in current use and include BDMax, FilmArray, Luminex, Prodesse, and Verigene. Gastrointestinal CIDT panels are advantageous over culture-based diagnostics for several reasons. First, the gastrointestinal panels are designed to detect multiple pathogens in a single assay, in some cases (e.g., FilmArray) up to 22 microbes. Commercially available assays differ in their microbial targets detecting distinct subsets of bacteria, viruses and/or parasites. Second, the various assays have different TAT – most ranging from 1 to 4 h – though even the most time-intensive assays (8 h) are considerably faster than stool culturing which can take 2–5 days dependent on the microbe being cultured. Third, multiplex panels retain the ability to detect polymicrobial infections, which is particularly relevant in clinical settings. Lastly, multiplex panels are highly sensitive (90–100%) and specific (98%) (Buchan et al., 2013; Navidad et al., 2013; Onori et al., 2014; Buss et al., 2015; Harrington et al., 2015).

NEXT-GENERATION SEQUENCING AS A DIAGNOSTIC ASSAY

Whole Genome Sequencing

Next-generation sequencing and other high-throughput laboratory techniques are circumventing laborious testing by active replacement or complement to traditional microbiological and molecular tests for identifying, typing and characterizing pathogens. An increase in read-length, output and quality of short-read NGS technologies has made it possible to apply WGS as a genomic surveillance system of foodborne diseases and outbreak management (Köser et al., 2012). In most cases, WGS provides higher discriminatory power than the combined effort of multiple conventional typing assays such as PFGE, MLST, MLVA, phage typing, virulence typing, and AST. Moreover, WGS is more reliable in a shorter time period and can be performed in a single comprehensive procedure thus allowing for rapid and sensitive pathogen identification and similarly for reporting to the food industry and government responsible for public health decision-making (Ashton et al., 2016; Jackson et al., 2016).

For some clonal microbes, molecular typing methods have proven unable to accurately discriminate genetically distinct isolates. As an example, in past outbreak investigations, highly clonal *Salmonella enterica* serovars such as *S. enteritidis* have been challenging to resolve using PFGE (the gold standard typing method). Taylor et al. (2015) reviewed seven epidemiologically supported clusters and revealed that SNV analysis was able to accurately discriminate cluster isolates from sporadic and suspect samples with high epidemiological concordance. PFGE results would have suggested that distinct outbreak isolates with no epidemiological link might have originated from the same source. The low resolution of PFGE was also exemplified in the 2008 listeriosis outbreak whereby suspected cases emerged with a mixture of two distinct but closely related PFGE patterns (Gilmour et al., 2010); WGS analysis revealed that isolates from both PFGE patterns were genetically similar apart from a large prophage responsible for the 40 kb band shift difference seen in the PFGE patterns of suspected isolates. WGS was only a research tool in those early days of public health genomics; however, its use in this large-scale outbreak was able to confirm that outbreak strains subtyped by either PFGE pattern were genetically similar apart from the insertion of a large prophage. The inclusion or exclusion of clinical strains associated with the outbreak and the resulting case definition can alter the case epidemiology and food safety investigation efforts.

Several NGS sequencing platforms are available; each with their own advantages and disadvantages differing in sequencing time, read length, cost and others, reviewed elsewhere (Mardis, 2017). At present, WGS, though still dependent on the presence of isolates, represents one of the predominant investigative tools to rapidly and accurately identify microbes, perform subtyping, cluster epidemiologically relevant isolates in outbreak investigations and detect AMR genes and virulence profiles. The use of WGS has also proven exceedingly useful in retrospective outbreak investigations. In 2013, PulseNet USA incorporated

WGS as a surveillance tool for all *Listeria monocytogenes* isolates (Jackson et al., 2016) and is poised to begin WGS of *Campylobacter*, STEC and *Salmonella*; PulseNet Canada is similarly performing routine WGS of all *L. monocytogenes* and intends to begin sequencing all *Salmonella* isolates routinely in 2017. Moreover, PulseNet Canada also performs WGS on select cluster or outbreak investigations of STEC and other organisms.

Diagnostic Metagenomics and Comparable Detection Techniques

Diagnostic approaches including culture and non-NGS CIDTs such as PCR or serology continue to represent the gold standard for infectious disease diagnostics. Each of these methods, however, is disadvantaged in that they represent a targeted detection methodology and hence *a priori* knowledge or hypotheses to identify the etiological agent in the sample are required. It has been suggested that the limited capacity of conventional diagnostics including culture and non-NGS CIDTs are partly responsible for failing to detect an etiological culprit in a considerable number of cases (Denno et al., 2012). High-throughput targeted-amplicon and shotgun metagenomics sequencing methods circumvent this limitation via broad-range detection of either a subset of microbes (targeted-amplicon) or all microbes (shotgun metagenomics). Further, diagnostic shotgun metagenomics offers an added advantage with the possibility to identify previously uncharacterized microbes or emergent and novel pathogens (Forster et al., 2016).

The dramatic increase of microbiome research in the last decade is effectively driven by the widespread usage of high-throughput DNA sequencing and the associated decrease in sequencing cost. In the context of the microbiome, NGS allows for the complete description of all genomic content of microbial communities (e.g., bacterial, viral, eukaryotic microbes) in a technique referred to as shotgun metagenomics (Sharpton, 2014). Though most microbiome studies are fundamentally designed to describe commensal populations (The NIH HMP Working Group, 2009) or alternatively, to investigate dysbiosis in distinct human body compartments affected by disease (Forbes et al., 2016), or to determine the efficacy of prebiotics (Alfa et al., 2017) and others, metagenomics can similarly be utilized for the identification of pathogens in clinical (Mongkolrattanothai et al., 2017) or food (Aw et al., 2016) samples. Like other molecular or serological CIDTs, the detection of microbes in metagenomics is independent of culture (Schloss and Handelsman, 2005) in contrast to WGS, which is also reliant on a pure culture (Hasman et al., 2014). Metagenomics and comparable detection techniques generate immense quantities of large-scale sequence data thus bioinformatics and/or computational approaches are required to assign sequences to particular microbes, microbial functions or other descriptors of relevance (Sharpton, 2014).

The majority of microbiome studies to date have focused specifically on the bacterial portion of microbiomes rather than characterizing the microbial communities across all domains of life in addition to viruses (shotgun metagenomics). As such, these analyses are often performed by high-throughput targeted-amplicon sequencing of a universal phylogenetically informative

genetic marker; the 16S rRNA gene is most commonly used though other markers are similarly able to discriminate between prokaryotes including *cpn60* (Schellenberg et al., 2016), *rpoB* (Case et al., 2007), 23S rRNA (Anthony et al., 2000), and others. Universal targets for eukaryotic organisms include the 18S rRNA gene (Kataoka et al., 2016) or alternatively, for specifically characterizing fungal populations the ITS is frequently used (Argenio et al., 2016).

While the science community has persistently been using the term metagenomics interchangeably to describe both high-throughput targeted-amplicon and shotgun metagenomics studies to profile microbial populations, it is important to make a clear distinction between both methodologies. Targeted-amplicon and shotgun metagenomics sequencing each present with advantages and disadvantages (Table 1). Features differentiating the two methods include targeted microbial community, associated costs and computational and technical expertise requirements among others. Thus, the decision to use one approach versus the other should be made with careful consideration and will be highly dependent on the research and/or diagnostic goals and hypotheses in question (Table 2). In this regard, targeted-amplicon assays are more apt for describing a specific group of microbes (e.g., bacteria) whereas a shotgun metagenomics approach is more suitable for characterizing the entirety of microbial DNA in a given sample limited by sequencing technology used and sample matrix with high host DNA-containing material.

Metagenomics and Comparable Detection Techniques in Food Safety

Culture-based techniques are still considered the gold standard in the food industry, including sectors such as business operators, government regulatory agencies and national or international compliance testing. Countries such as Canada have policies in place that require quantitative (Pagotto et al., 2011a) or qualitative (Pagotto et al., 2011b) approaches to determine viable pathogens such as *L. monocytogenes*¹. This is attributed to how foods are categorized based on the ability to support (or not), the priority assigned to high-risk foods and the ability of the pathogen to survive during the shelf life of the food¹. Moreover, the target pathogen is often in such low numbers that in the presence of the background microbiota, enrichment is required – and culturing is the most effective strategy (Gill, 2017). While a molecular fingerprint *per se* is not currently required for compliance activity, it is invaluable to epidemiological investigations and for source attribution. As such, current approaches necessitate a physical isolate in order to generate molecular fingerprints. Further, it is important to note that more than one “molecular type” may be implicated in an outbreak and so, culture-dependent technologies continue to have an important role. This was shown in the *Listeria cantaloupe* outbreak where multiple serovars and five different molecular subtypes were found (McCollum et al., 2013). It is anticipated that bioinformatics will help play a role in being able to differentiate

¹http://www.hc-sc.gc.ca/fn-an/legislation/pol/policy_listeria_monocytogenes_2011-eng.php

TABLE 1 | Summary of the advantages and disadvantages to each high-throughput sequencing approach for unbiased detection.

	Targeted-amplicon sequencing		Shotgun metagenomics sequencing	
	Advantages	Disadvantages	Advantages	Disadvantages
Microbial target(s) of interest	<ul style="list-style-type: none"> Target is specific to a particular microbial group (e.g., 16S rRNA common for bacteria, archaea, 18S rRNA for eukaryotes, ITS for fungi, RdRP for RNA viruses). 	Requires <i>a priori</i> knowledge for microbial group target.	<ul style="list-style-type: none"> Can sequence all DNA in a given sample (e.g., bacteria, archaea, eukaryotes, parasites, and viruses). 	Virome assays require complex sample and nucleic acid work-ups.
Abundance profiling	<p>Can use relative abundance changes to compare microbiomes across different samples or treatments.</p> <ul style="list-style-type: none"> Can capture abundance of rare taxa provided that sequencing depth is sufficient. 	Universal target chosen may be present in varying copy numbers across different taxa (e.g., 16S rRNA). PCR amplification bias, primer bias and errors.	<p>Universal markers can be inferred from metagenomics datasets.</p>	<ul style="list-style-type: none"> High abundance of host DNA can make it challenging to sequence low abundance microbial DNA.
Taxonomic assignment	<ul style="list-style-type: none"> Relatively easy to taxonomically classify sequences using a variety of validated tools and curated databases. 	<ul style="list-style-type: none"> Databases can be self-limiting and have the potential to exclude novel microbes. Universal targets within microbial groups can give variable taxonomic classifications. Taxonomic resolution variable – species level identification should be interpreted with caution. 	<ul style="list-style-type: none"> Plethora of software using phylogenetically informative gene markers. 	<ul style="list-style-type: none"> Low abundance taxa difficult to identify. Can be difficult to accurately bin each sequence to a genome. High proportion of taxonomically uninformative sequences are discarded. Availability and access to comprehensive and curated databases across all microbial groups limited.
Cost	<ul style="list-style-type: none"> Low cost Can be carried out on most bench-top sequencers and sequencing platforms. 		<ul style="list-style-type: none"> Can be carried out on most bench-top sequencers and sequencing platforms. 	<ul style="list-style-type: none"> Can be cost prohibitive depending on the sequencing depth, sample type, and microbe(s) of interest. If high host DNA is expected or interest is in the low-abundance microbes or rare taxa, use of a higher throughput sequencer (Illumina HiSeq), may be required. High performance computing environment absolutely necessary.
Computational requirements	<ul style="list-style-type: none"> Most analysis steps can be carried out on a modern desktop. 	<ul style="list-style-type: none"> Large datasets (high sample number and/or sequencing coverage) may require access to a high performance computing cluster dependent on analytical pipeline chosen. 	<ul style="list-style-type: none"> Cloud computing services are available for metagenomics data analysis for those without access to a high performance computing cluster. 	<ul style="list-style-type: none"> Cloud computing – potentially cost-prohibitive and might not have all available pipelines and/or software. Data privacy and sensitivity may prohibit the use of commercial cloud computing services. High technical expertise required.
Technical expertise	<ul style="list-style-type: none"> Moderate to high technical expertise is required depending on the analytical pipeline chosen. 			

TABLE 2 | Overview of appropriate usage for each unbiased high-throughput sequencing approach.

Study goals/purpose	Suggested sequencing approach
Characterization of a particular microbial group (excluding viruses) in sample(s)	High-throughput targeted-amplicon sequencing; utilize shotgun metagenomics sequencing if interested in high taxonomic resolution above genus level.
Characterization of all microbial DNA in sample(s)	Metagenomics shotgun sequencing.
Pathogen detection	Dependent on the sample:
	<ul style="list-style-type: none"> • If the etiological agent is suspected to be of viral origin a shotgun metagenomics approach is warranted. • If the sample type contains a high host DNA load (e.g., blood) should consider a targeted-amplicon or deep shotgun metagenomics sequencing approach. The latter may be cost prohibitive and require access to a high-throughput sequencer (e.g., Illumina HiSeq). • Low biomass samples (e.g., BAL/CSF), might require a targeted-amplicon sequencing approach initially. Shotgun metagenomics sequencing may not be able to sequence the infectious agent adequately (e.g., only a few sequences produced which may only yield a confounding signal).
Functional profiling	Functional profiles can be inferred with a targeted-amplicon sequencing approach, however, results should be interpreted with caution due to the limitations of inferring gene function with universal targets. A shotgun metagenomics approach would yield more appropriate and reliable conclusions.
SNV or clonal isolate detection studies	Shotgun metagenomics sequencing.
Novel microbial identification and characterization	Targeted-amplicon sequencing relies on curated databases of known microbes and may not be able to adequately analyze novel microbes in an unbiased technique. Shotgun metagenomics would be recommended.

BAL, bronchoalveolar lavage; CSF, cerebrospinal fluid; SNV, single nucleotide variant.

the different molecular types present in a single food item (unlike a clinical specimen that tends to be contaminated with a single type). Molecular techniques that are currently used to compare isolates during an outbreak investigation, such as PFGE, ribotyping and gene-specific PCR, require an isolate as a starting point (Ronholm et al., 2016); these procedures would generally take approximately 7 days (**Figure 3**). If however, a high-sensitivity metagenomics methodology, capable of reliably detecting foodborne pathogens in samples with high-levels of background microbiota were developed, the time for outbreak recognition, causative agent sub-typing, secondary analyses for virulence and AMR genes and source attribution would be reduced (**Figure 3**). However, even after suitable techniques are developed, widespread use of CIDT as a replacement for culture-based techniques will still face significant challenges in food regulation and industry. International regulatory agencies with stakes in food safety such as Health Canada, US FDA, French agency for Food, Environmental and Occupational Health Safety (ANSES) are still in the initial phases of accepting WGS results as a replacement for more traditional techniques for outbreak investigations². Some countries though, are much further along in this process than others (Allard et al., 2016). Using Canada and *Listeria* as an example, verification controls in processing environments is directly tied to consumer risk should the food be contaminated (Pagotto et al., 2011a). The ecology of *Listeria* in ready-to-eat products with respect to foods capable (or not) of supporting growth (or survival) may be elucidated through the use of WGS. Metagenomics has been used to help determine aspects related to enrichment (Ottesen et al., 2016); interference to even the ability of only detecting a single species when more than one may be present (Ottesen et al., 2016). A relatively

recent study investigated the efficacy of NGS techniques to detect food pathogens (Leonard et al., 2015). An FDA culture-based protocol was performed on spinach spiked with STEC in addition to shotgun metagenomics sequencing. In particular, the study aimed to address limits of detection, sensitivity and specificity levels. The authors reported an expected level of contamination (approximately 10 cfu/100 g) and were able to accurately detect strain-level and virulence information within 8 h of enrichment at a sequencing depth of 10 million reads.

Next-generation sequencing-based approaches such as WGS rely on sequencing technologies, bioinformatics pipelines (Lambert et al., 2015) and high-quality reference databases (Allard et al., 2016). Therefore, widespread acceptance of WGS results, as a complete replacement for traditional molecular techniques must likely precede the introduction of shotgun metagenomics or targeted-amplicon approaches in food outbreak investigations and compliance testing. There are, however, many instances where metagenomics are being used in the food industry for rapid screening and research, but for regulatory testing purposes and for outbreak investigations, presumptive positive results must still be culturally confirmed³.

Metagenomics is currently being used in both the food industry and in research for many applications including: taxonomic profiling of microbiological food products and supplements, directing efforts to improve culture techniques, identification of non-culturable or fastidious pathogens and detection of co-contamination. Targeted-amplicon analyses for example, are well-suited to study, characterize and catalog changes in the bacterial populations that take place during fermentation reactions and long-term storage of traditional fermented foods such as soft cheese (Escobar-Zepeda et al., 2016),

²<http://www.fao.org/documents/card/en/c/61e44b34-b328-4239-b59c-a9e926e327b4/>

³<http://www.hc-sc.gc.ca/fn-an/res-rech/analy-meth/microbio/index-eng.php>

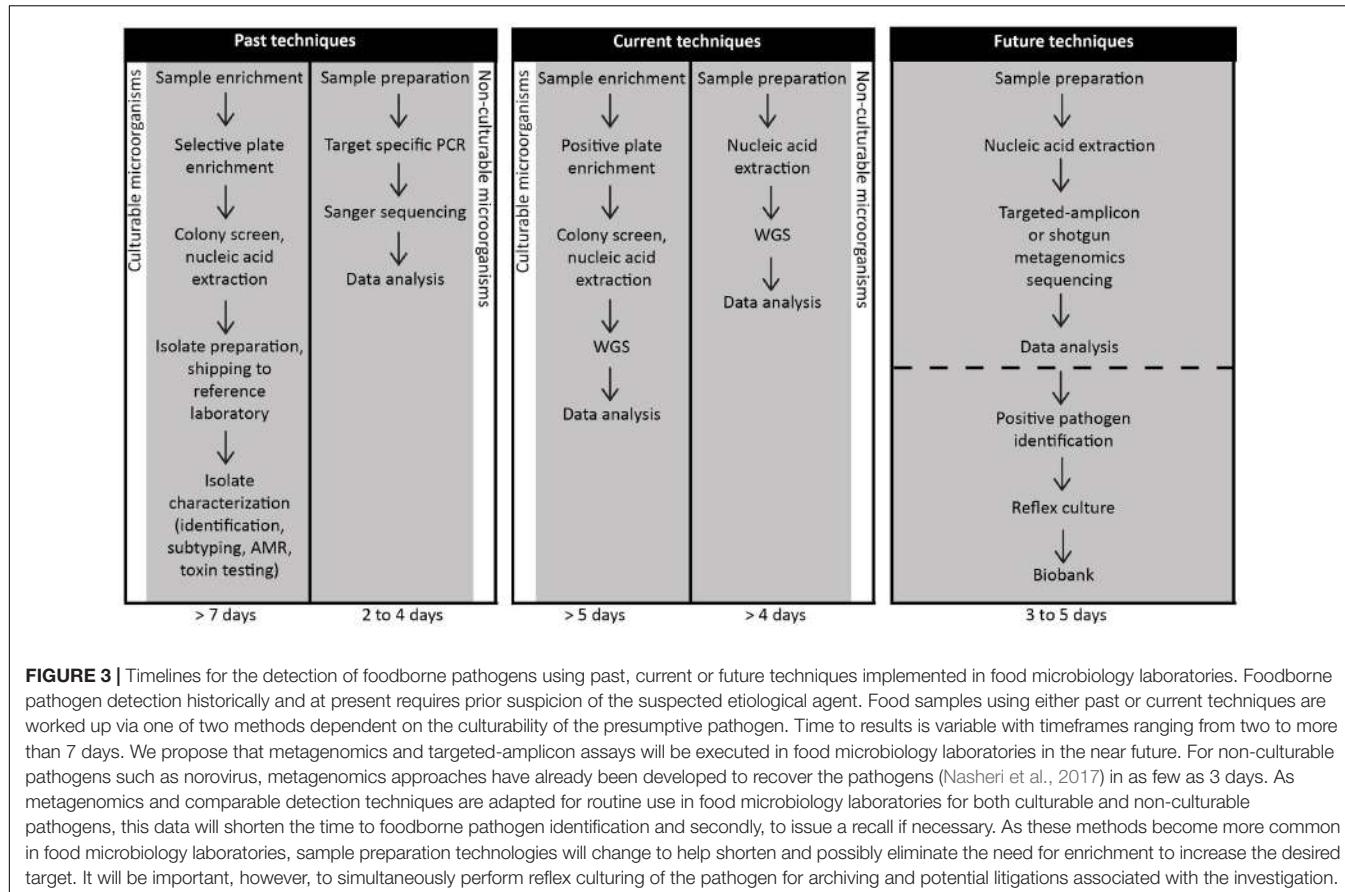


FIGURE 3 | Timelines for the detection of foodborne pathogens using past, current or future techniques implemented in food microbiology laboratories. Foodborne pathogen detection historically and at present requires prior suspicion of the suspected etiological agent. Food samples using either past or current techniques are worked up via one of two methods dependent on the culturability of the presumptive pathogen. Time to results is variable with timeframes ranging from two to more than 7 days. We propose that metagenomics and targeted-amplicon assays will be executed in food microbiology laboratories in the near future. For non-culturable pathogens such as norovirus, metagenomics approaches have already been developed to recover the pathogens (Nasheri et al., 2017) in as few as 3 days. As metagenomics and comparable detection techniques are adapted for routine use in food microbiology laboratories for both culturable and non-culturable pathogens, this data will shorten the time to foodborne pathogen identification and secondly, to issue a recall if necessary. As these methods become more common in food microbiology laboratories, sample preparation technologies will change to help shorten and possibly eliminate the need for enrichment to increase the desired target. It will be important, however, to simultaneously perform reflex culturing of the pathogen for archiving and potential litigations associated with the investigation.

kimchi (Jung et al., 2011; Park et al., 2012), and kefir (Leite et al., 2012; Korsak et al., 2015). Targeted-amplicon analyses are also suited to evaluate the composition of probiotic supplements (Bergholz et al., 2014; Morovic et al., 2016). This technique, when used to evaluate probiotic supplements, could be of particular interest to regulatory bodies, since based on preliminary work, many of these products contain neither the number nor species of probiotic bacteria claimed on the label (Morovic et al., 2016).

Targeted-amplicon and shotgun metagenomics have also both been used to improve culture-based enrichment techniques by allowing for detailed characterization of population dynamics of the background microbiota during enrichment (Gorski, 2012; Ottesen et al., 2013, 2016; Jarvis et al., 2015; Zilelidou et al., 2016). When food items are added to a non-selective media, microbes from the background microbiota can co-enrich with pathogens interfering with detection and recovery of pathogens. For microbes from the microbial population closely related to the pathogen this can also be true of differential or selective media as well. As an example, tomatoes are well-known to be frequently implicated as the source of human *Salmonella* infections, however, isolating *Salmonella* from the tomato phyllosphere using a culture-dependent method has proven challenging (Ottesen et al., 2013). Shotgun metagenomics sequencing of enrichment media revealed that *Paenibacillus* spp.

readily outcompetes and may even kill *Salmonella* during the enrichment phase suggesting that an alternate enrichment media be used when investigating tomato contamination (Ottesen et al., 2013). A similar enrichment problem occurs with cilantro, where traditional pre-enrichment steps encourage the growth of Gram-positive Firmicutes instead of Proteobacteria (*Salmonella*), suggesting that an alternate pre-enrichment media should be developed for cilantro testing (Jarvis et al., 2015). The usefulness of targeted-amplicon and shotgun metagenomics sequencing in informing enrichment strategies has similarly been demonstrated in *L. monocytogenes* (Ottesen et al., 2016) and *E. coli* (Margot et al., 2016).

The use of CIDTs are much more readily acceptable for outbreak delineation and regulatory testing for non-culturable or fastidious foodborne pathogens such as viruses (Vinjé, 2014) and parasites (Buss et al., 2013). For example, noroviruses, which until very recently were non-cultivable (Ettayebi et al., 2016), are currently detected by Sanger sequencing of partial regions of the polymerase and capsid sequences, directly from clinical samples (Vinjé, 2014). However, the same techniques are rarely successful on food samples since the viral titer in the food is generally too low (Gentry-Shields and Jaykus, 2015; Figure 3). Sanger sequencing of norovirus requires several PCR rounds to obtain usable information, and still only yields partial genome sequences (Verhoef et al., 2012). To successfully

delineate outbreaks with 100% specificity, a full capsid sequence is required (Nasheri et al., 2017). Currently, several WGS workflows exist for norovirus analysis, including the sequencing of several overlapping PCR fragments (Kundu et al., 2013; Won et al., 2013), target enrichment of the norovirus genome using custom designed RNA baits (Brown et al., 2016) and metatranscriptomics approaches (Bavelaar et al., 2015; Nasheri et al., 2017; **Figure 3**). The metatranscriptomics approach has the advantage of being rapid and further, a *de novo* assembly can be obtained. This approach is thorough and additionally, allows for the presence of other co-infections to be detected from the same sequence data (Nasheri et al., 2017). Metagenomics sequencing has also been used to investigate various food products for the presence of potential or emerging pathogens in the absence of a defined outbreak (Kawai et al., 2012; Aw et al., 2016). A shotgun metagenomics approach has recently identified several human and animal viruses in fresh produce (Aw et al., 2016), and it can be expected that this type of analysis will be repeated for other food products to provide greater insight into the scope of viral contamination of the food supply. Non-culturable parasites present similar challenges to non-culturable viruses; shotgun metagenomics has also provided a suitable solution for detecting parasitic contamination. Between 2008 and 2010, Japan experienced multiple outbreaks of gastroenteritis that included more than 1300 cases caused by an unknown etiological agent. Shotgun metagenomics was used to demonstrate that a parasite (*Kudoa septempunctata*) was the likely cause of the outbreak (Kawai et al., 2012).

The current regulatory dilemma is that it is impossible to tell directly from the presence of DNA if an organism is viable or not, and although DNA degrades over time, false positive rates of pathogen identification due to the detection of naked DNA have the potential to be quite high. This is where food differs from clinical samples. If pathogen DNA is detected in clinical samples, there is a very high-probability that the pathogen is alive and replicating in the patient. However, finding pathogen DNA in foods has quite a different interpretation. A well-known example is *Listeria* in smoked fish, where its DNA can be present due to dead cells but molecular methods alone would generate incorrect compliance activity based on a PCR-based method (Gambarin et al., 2012; Law et al., 2015). In addition, food-processing techniques (such as thermal, high-pressure, radiation exposure, and others) are known to kill bacteria, but leave detectable pathogen DNA in the food matrices. This problem with false-positive detection is a significant hurdle to overcome before metagenomics approaches will be useful in outbreak delineation or compliance testing in the food industry (Law et al., 2014). Moreover, when secondary analyses including virulence or AMR gene detection are applied to the metagenomics data, and particularly if the genes are known to be mobile, there is a current lack of bioinformatics pipelines available to accurately predict if the virulence or AMR genes belong to the pathogen, or rather to the background microbiota. It has been suggested that the alternative use of mRNA could contribute to solving this problem, although, its ability to persist in various samples has also been implied (Leimena et al., 2013).

Diagnostic Case Studies Revealing the Usefulness of Metagenomics and Comparable Techniques

Numerous studies have investigated the capacity of diagnostic metagenomics for infectious diseases from clinical specimens; these studies have proven the methodology clinically useful from an analytical perspective and secondly, in an appropriate timeframe that can nevertheless yield positive patient outcomes (Nakamura et al., 2008; Wilson et al., 2014; Mongkolrattanothai et al., 2017). Both high-throughput targeted-amplicon and shotgun metagenomics sequencing have been performed on biological specimens for unknown pathogen detection either for research purposes or when all previous diagnostic tests have proven inconclusive. Importantly, the use of both high-throughput sequencing strategies to determine etiological causation is highly experimental due to inherent limitations (discussed in upcoming section) and typically only performed on select cases in diagnostic settings. Preceding widespread adoption and implementation of these methodologies, case studies are needed to evaluate concordance and compatibility with conventional diagnostic methods. Interpretation and reporting guidelines, in addition to criteria are likewise required prior to real-time use of this technology.

From a diagnostics perspective, metagenomics sequencing was first employed to detect human-associated viruses (rather than bacteria), which is unsurprising given the challenges associated with culturing most viruses. Throughout the last decade, viral metagenomics (viroomics) has been used as a tool to diagnose cases of acute gastroenteritis and further, to determine etiological causation retrospectively in gastroenteritis outbreaks (Finkbeiner et al., 2009; Smits et al., 2014). This technique has proven particularly useful in determining causative agents whereby culture yielded negative results. It is hypothesized that novel viruses may account for a fraction of idiopathic gastroenteritis cases especially given that a large portion of global viral diversity has yet to be discovered (Anthony et al., 2013); hence methodologies incorporating “unbiased” and systematic high-throughput sequencing have the capacity for novel virus detection and discovery, which may have otherwise evaded routine diagnostic testing. In this regard, an early study aimed to characterize the viral populations present in pediatric diarrheal specimens (Finkbeiner et al., 2008). Interestingly, this study adapted a “micro-mass sequencing” approach. This included a minimal sample quantity (<100 mg stool), minimal sample purification and minimal sequencing (e.g., 384 reads per sample). Known enteric viruses including rotaviruses, caliciviruses, astroviruses, and adenoviruses were detected in addition to several sequences from at least nine putative novel viruses. Recently, a shotgun metagenomics approach was applied retrospectively to detect novel and known viruses associated with gastroenteritis outbreaks (Moore et al., 2015). Though the study failed to identify any novel pathogens, eight viruses, and one parasite were detected. The authors’ concluded that metagenomics could be useful to detect pathogens whereby routine testing has failed.

An early study tested the capability of shotgun metagenomics sequencing to detect bacterial pathogens during a case of acute gastroenteritis following culture based diagnostics where no candidate pathogens were identified (Nakamura et al., 2008). This case reported 156 *Campylobacter jejuni* sequences from a stool sample obtained when the patient was symptomatic versus no sequences identified in a sample obtained 3 months post-illness. This investigation represents one of the first proof-of-concept studies to be conducted in a clinical diagnostic setting.

A relatively recent case study emphasizes the clinical relevance of metagenomics as a diagnostic assay (Wilson et al., 2014). Briefly, a 14-year-old boy presented to a medical facility three times with complaints of fever and headache that ultimately advanced to hydrocephalus and status epilepticus. Though diagnostic workup was inconclusive on several occasions, an MRI eventually revealed encephalitis-like indications. CSF and serum specimens were subjected to shotgun metagenomics sequencing. Within 48 h of specimen receipt, bioinformatics analysis of CSF revealed a high abundance of sequences with homology to the *Leptospiraceae* family and mapped most closely to the pathogenic *Leptospira borgpetersenii* genome. The patient was therefore treated for neuroleptospirosis with intravenous penicillin G and markedly improved. In this context, treatment was initiated on the basis of metagenomics evidence and congruent clinical presentation prior to the completion of validated confirmatory testing. Five months post-treatment, targeted PCR and Sanger sequencing identified *L. santarosai* as the infectious agent. This case highlights various advantages and limitations associated with diagnostic metagenomics that will therefore be discussed in the following section.

Diagnosis of brucellosis historically has been a challenge as this illness presents with a continuum of clinical manifestations. In addition, diagnosis is further limited by current diagnostics via their inadequate sensitivity and specificity (Mongkolrattanothai et al., 2017). A recent report described a case whereby a shotgun metagenomics analysis of the patient's CSF was used to provide an accurate diagnosis (via *Brucella* spp. detection) and directed towards appropriate antibiotic therapy ultimately leading to a favorable patient outcome (Mongkolrattanothai et al., 2017). This study led to the development of a validated diagnostic assay in the CLIA-certified University of California, San Francisco clinical microbiology laboratory⁴.

A recent study aimed to assess the concordance between various diagnostic (PCR, qPCR) and NGS techniques (16S rRNA targeted-amplicon and shotgun metagenomics) for *C. difficile* infection (Zhou et al., 2016). Intriguingly, of PCR and qPCR positive *C. difficile* samples, this microbe was detected in 90.9% of samples via 16S rRNA analysis versus 86.3% with shotgun metagenomics. Moreover, *C. difficile* was co-detected with several known enteric pathogens such as norovirus and sapovirus which adds further credence to the efficacy of NGS CIDTs to firstly, detect foodborne pathogens in an unbiased technique and secondly, to detect

polymicrobial infections. This study reflects current limitations of metagenomics techniques in that their sensitivity is still relatively low.

We mentioned previously that high-throughput targeted-amplicon sequencing is at the moment largely research based. A recent study described the usefulness of 16S rRNA targeted-amplicon sequencing in an exploratory setting (Almonacid et al., 2016). This study utilized 46 bacteria and archaea encompassing 15 genera and 31 species of microbes in the human gastrointestinal tract selected on the basis of clinical relevance including pathogenic, commensal, and probiotic prokaryotes. A bioinformatics annotation pipeline specifically tailored to have high prediction performance was applied. For example, taxonomies were reported based on 100% identity over the entire 16S rRNA V4 region and curated databases were procured for each taxon using various optimizing parameters such as sensitivity, specificity, precision and a negative predictive value. Applying a threshold of 90% for each parameter, 28 of 46 targets could be detected. Microbiome specimens from a cohort of 897 healthy persons were used to define a reference range that could be used to establish clinically applicable relative abundances for each target. The authors concluded that their assay accurately identified and quantitated all targets and known pathogens from each sample (clinical or synthetic). Thus, assays such as this may facilitate improvements to patient diagnosis, treatment, monitoring, and epidemiological study.

The use of diagnostic metagenomics is less established in the framework of parasitic infections; however, studies are beginning to investigate their role (or merely, presence) in the human gastrointestinal tract, in addition to elucidating genomic epidemiology. Moore et al. (2015) also detected a parasitic candidate – *Dientamoeba fragilis*. While the role of *D. fragilis* in causing gastroenteritis remains unclear it is suggested that in the absence of other enteric pathogens that this parasite could be considered a causative agent (Barratt et al., 2011). This study also revealed an interesting phenomenon: while *D. fragilis* was detected in 10.9% of undiagnosed outbreak samples, a higher frequency was reported (44%) in pediatric samples. A more recent study applied several diagnostic techniques (metagenomics, microscopy, and multiplex PCR) to four diarrheal samples as a means to detect multiple pathogens simultaneously (Schneeberger et al., 2016). Metagenomics detected 8–11 plausible enteric pathogens in all samples. Specifically with bacterial pathogens, diagnostic agreement between PCR and metagenomics was high though metagenomics did identify several bacteria not detected by PCR. Perhaps more interesting, however, was the finding that microscopy could detect some helminth and protozoan infections that metagenomics could not, again reflecting sensitivity issues inherent of metagenomics techniques. Parasitic metagenomics has also shown effective in the investigation of *Cyclospora cayetanensis* – a coccidian parasite responsible for several food and waterborne outbreaks worldwide (Cinar et al., 2015) and in a case study of Malayan filariasis caused by *Brugia malayi* (Gao et al., 2016).

⁴<https://www.ciapm.org/project/precision-diagnosis-acute-infectious-diseases>

Technical Challenges of Metagenomics Sequencing and Analysis

The diagnostic applicability of clinical metagenomics presents with its own shortcomings and hence is currently limiting the widespread use of this approach in clinical and food microbiology laboratories. In this regard, the aforementioned neuroleptospirosis case highlights many key points with respect to the utility of metagenomics in a diagnostic setting. First, Wilson et al. (2014) reported that a diagnosis would not have been possible via conventional assays. As is the case with many infections, differential diagnoses can be extensive; consequently, cultures, serology, and pathogen-specific PCR can be difficult in determining etiological causation in cases of this magnitude. These difficulties would particularly benefit from the use of unbiased diagnostic methods. Second, reference databases (e.g., NCBI, UniProt for protein and SILVA for ribosomal RNA) are a cornerstone of NGS-based diagnostics, however, the level of curation within these databases is variable. In the absence of adequately curated databases, accurate diagnoses will not be possible and could be detrimental to patient outcome should a misdiagnosis occur. In the neuroleptospirosis case, shotgun metagenomics and bioinformatics analyses utilizing the NCBI reference database identified *L. borgpetersenii*; confirmatory testing revealed the causative agent to be *L. santarosai*. In this context, query sequence homology to a reference sequence should not be misconstrued as causation. Third, the exceptionally short TAT led to appropriate therapeutic management and thus clinically actionable results, which corresponds to our next discussion point – perhaps most importantly is the developmental need for analysis tools and bioinformatics pipelines modernized for the requirements of diagnostic laboratories. For example, tools for pathogen detection must provide analytical results while concurrently minimizing analysis time to allow for reasonable diagnostic TAT. Though several analyses software or platforms for microbial identification are available (Segata et al., 2012; Naccache et al., 2014; Wood and Salzberg, 2014; Flygare et al., 2016), appropriate use typically requires bioinformatics or computational biology expertise often not available in diagnostic laboratories. One such computational analysis pipeline recently developed, SURPI (Naccache et al., 2014), is designed for unbiased pathogen detection from shotgun metagenomics sequencing data of clinical specimens applicable to infectious disease diagnostics, public health surveillance and outbreak analysis. Currently, practical utilization of this technology is hindered by several computational and technical challenges of analyzing data accurately and in a clinically actionable timeframe. Specific requirements include access to high performance computing, as well as laboratory technicians that are cross-trained in bioinformatics (ability to generate data) and in biology (provide interpretation of the data). There is an important need to develop clinically useful pipelines such that metagenomics can be implemented as a diagnostic tool.

Not only are challenges associated with the analytical performance, but sample preparation and sequencing methodologies must also be standardized. Due to the infancy of this field, standardizations are still in progress and various

ring trials and proficiency tests are underway to standardize various aspects pertaining to studies of this nature. For further information we direct readers to the following articles (Lesko et al., 2016; Mellmann et al., 2017). As outlined in **Tables 1, 2**, the choice of sequencing technique applied is largely dependent on the research or diagnostic goals while also considering advantages or disadvantages of each method. Other factors that can affect pathogen identification include as previously discussed, pre-enrichment steps, quality of DNA extraction, library preparation and chemistry. Furthermore, we are faced with questions pertaining to how to work with contaminants or host DNA, low biomass specimens, defining ideal read depths for particular biological specimens or food products and lastly, defining sequence number thresholds to confidently assign pathogenic etiology.

The presence of contaminants or host DNA poses serious challenges to metagenomics data analysis whereby in many instances, true (e.g., microbial) metagenomics data is overwhelmed by the abundance of host DNA. Considerations should be given to samples where eukaryotic cell content is expected to be high (e.g., biopsy). Similarly, variable host DNA content will be found in samples such as stool dependent on the consistency and quality (e.g., healthy, watery, or bloody). Studies have attempted to address the issue of contaminants (Salter et al., 2014) and host DNA (Hasan et al., 2016) with variable success. The primary concern, however, are methodologies surrounding removal of these products and their plausible effect on microbial DNA abundance or quality in a given sample.

While we frequently refer to metagenomics as an unbiased methodology, there are biases that need to be considered to ensure that all microbes of interest are captured adequately in the sequencing output. As an example, interrogating the DNA or RNA virome using a metagenomics approach (e.g., viromics) requires complex wet-laboratory procedures that are considerably distinct from that of a sample workup for a general metagenomics assay of ‘all’ microbes. Viruses have significantly smaller genomes as compared to prokaryotes or eukaryotes, thus their genomic content represents a microscopic fraction of total DNA or RNA in a given sample (Kleiner et al., 2015). Moreover, characteristics of viral particles are considerably diverse (e.g., size, density, structure, and others) and also include phages. In this context, an intricate series of sample and nucleic acid treatments must be conducted to ensure viral particles are amply interrogated with high sensitivity in a metagenomics assay (Thurber et al., 2009).

Within a given specimen, a broad range of taxon abundances are observed. Though shotgun metagenomics can produce near-complete whole-genome coverage of highly abundant microbes, etiological agents are not always among the most abundant. For example, while the infective dose (in colony forming units) is 10^7 – 10^9 for enterotoxigenic *E. coli*, *L. monocytogenes* is 10^3 and *Salmonella* spp. is only 4–50 (Todd et al., 2008). Achieving sufficient sensitivity in shotgun metagenomics studies has proven to be a challenge, particularly with respect to low abundance microbes and gene level identification. For example, data mining for AMR, toxin or virulence genes has proven difficult. This was

well-established in a retrospective metagenomics study that aimed to investigate an outbreak of STEC O104:H4 (Loman et al., 2013). The authors' reported that the outbreak strain was recovered to near completeness (full genome breadth) from 10 samples (of 45) at >10-fold coverage and from 26 samples at >1-fold coverage. The shiga-toxin gene, however, which was identified in 100% of samples via culture-based methods was only detected in 27 of 40 (67%) STEC-positive samples. This study highlights some of the challenges of metagenomics related to gene level sensitivity. Further examination, particularly into the acceptable sequencing depth to ensure high sensitivity and specificity is warranted.

Metagenomics has already proven an effective complement to conventional diagnostics in complex cases and outbreaks (Loman et al., 2013; Ruppé et al., 2016; Langelier et al., 2017) though much remains to be elucidated prior to the widespread adoption of metagenomics in diagnostic and public health laboratories. Therefore, while high-throughput sequencing of biological specimens has a promising future, its utility is unlikely to become a standard and approved CIDT method in the imminent future.

INFLUENCE OF CULTURE-INDEPENDENT DIAGNOSTIC TESTING ON HUMAN DISEASE SURVEILLANCE

Pathogen Surveillance and Subtyping

Metagenomics sequencing of a clinical or environmental sample could offer a universal test for pathogen detection, clinical diagnosis, as well as a subtyping test for routine surveillance activities. Surveillance for foodborne infections historically has required the culturing of an isolate in order to perform the appropriate genotypic or phenotypic characterization, yielding a phenotype or genetic fingerprint capable of making detecting and resolving outbreaks possible. By tracking foodborne pathogens along the farm to fork to clinical specimen continuum, it is possible to monitor trends over time, track which foods are capable or implicated in causing illnesses and detecting outbreaks.

Similar to trends occurring in infectious disease diagnostics, techniques for subtyping have enhanced the ability to distinguish between epidemiologically linked isolates from identical microbial species hence improving outbreak detection, surveillance and the overall understanding of microbial epidemiology. Molecular subtyping techniques in routine use include PFGE and MLVA; PulseNet Canada⁵ (national real-time surveillance and outbreak response) has relied on these for nearly two decades but is presently transitioning to the use of WGS. The other national surveillance systems dedicated to foodborne disease in Canada, the National Enteric Surveillance Program⁶ (weekly trend analysis at the species or serotype

level of bacterial, viral and parasitic enteric pathogens) and FoodNet Canada⁷ (sentinel site-based surveillance system measuring burden of illness, attribution studies and food safety policy recommendations) are also transitioning from molecular subtyping to WGS.

Genomic Epidemiology

Recent improvements of sequencing technologies and streamlined bioinformatics tools have not only advanced clinical diagnostics but are also transforming public health. WGS is increasingly implemented in epidemiological study, outbreak detection and surveillance of foodborne bacteria. Genomic epidemiology refers to the use of WGS to investigate epidemiological features. The listeriosis outbreak described above was the first application of WGS in during an active foodborne outbreak investigation (Gilmour et al., 2010). This study was seminal in bridging the gap between WGS and public health in real-time. Further, the initial study to utilize WGS for source attribution was performed during the 2009–2010 *S. enterica* serotype Montevideo outbreak (Lienau et al., 2011). The source was traced to red and black pepper that was used in the production of Italian-style spiced meats in a New England processing facility. WGS has since been executed in manifold outbreak and surveillance analyses.

Between 2010 and 2015, numerous severe illnesses associated with a complex multi-state listeriosis outbreak were reported and linked to two facilities of a large commercial ice cream producer as the source of *L. monocytogenes* (Jackson et al., 2016). This outbreak is highlighted here particularly due to the unusual length of the outbreak. Specifically, guidelines pertaining to listeriosis outbreak investigations have generally used a 120-day window (versus 16-days for other foodborne pathogens) for inclusion of suspected cases attributed to its psychrotrophic nature and viability in cured and processed food products with longer shelf lives. WGS routine implementation for all clinical, food and food processing environmental isolates, for current and retrospective cases led to strong evidence supporting a lengthy listeriosis outbreak. Traditional investigation guidelines and culture-based subtyping methods (e.g., PFGE) would not have been able to unequivocally link *Listeria* isolates to an outbreak cluster. The high discriminatory power of WGS combined with strong epidemiological evidence will inevitably lead to a higher proportion of detected and resolved outbreaks and a concomitant lower number of patients within each cluster, thus allowing for contaminated products to be removed from commerce more promptly.

Metagenomics is Capable of Providing Informative Subtyping Data

Advancements have clearly been made with respect to pathogen detection via the application of molecular diagnostic techniques such as PCR. These advancements are driven by the capacity to bypass the need for pathogen culture and isolation. Current surveillance methodologies (e.g., PFGE, MLVA, MLST, and WGS) however, are reliant upon the presence of isolates.

⁵<https://www.ncbi.nlm.nih.gov/PulseNet/index-eng.htm>

⁶<https://www.ncbi.nlm.nih.gov/NESP-PNSME/index-eng.htm>

⁷<http://www.phac-aspc.gc.ca/foodnetcanada/index-eng.php>

Techniques used in public health surveillance for disease tracking and subtyping therefore necessitate adaptation to the culture-independent trend.

Rapid infectious disease diagnostics would particularly benefit from the ability to directly subtype pathogens from complex clinical specimens. While subtyping methods such as targeted-amplicon sequencing, FISH (Splettstoesser et al., 2010) and repetitive element sequence-based PCR (Hahm et al., 2003) may be able to interrogate and subtype microbes, their ability to do so at an adequate taxonomic resolution renders these assays less efficient in differentiating between subtypes of a specific species, compared to gold standard methods such as PFGE. Metagenomics sequencing of clinical specimens represents a plausible epidemiological and subtyping tool. Hundreds to thousands of sequence reads for a particular species are generated through metagenomics sequencing thus potentially providing sufficient informative data for subtyping. Genome coverage of a given microbe, however, is difficult to predict and often the pathogen may not be the most abundant microbe in a given specimen due to the pathogen's infective dose. Also, some serogroups are unable to be typed, for example "O rough" STEC (Chattaway et al., 2016). In recent years, the widespread usage of research-based metagenomics has coincided with the development and application of a plethora of novel analysis techniques, some of which are well-suited to type (e.g., interrogate the microbe below species level taxonomic resolution) microbial strains (Zagordi et al., 2011; Hong et al., 2014; Ahn et al., 2015; Cleary et al., 2015; Sahl et al., 2015; Joseph et al., 2016). Numerous analysis software have been developed and similarly been shown to achieve adequate sensitivity and specificity for pathogen identification. As of yet, metagenomics to our knowledge has not been utilized in a communicable disease tracking perspective.

Changing Trends in Public Health Surveillance

Effective detection of outbreaks, particularly broadly disseminated outbreaks caused by the commercial distribution of contaminated foods is largely dependent on subtyping isolates from a sizeable proportion of cases. As diagnostic laboratories are in the process of shifting from diagnostic tests (culture-based or WGS) yielding isolates to CIDTs (molecular or serological), it is essential to maintain a system whereby subtyping can still be performed on a large percentage of positive specimens.

With changing diagnostic practices, several options to maintain the capability of foodborne surveillance, outbreak detection and source attribution are possible. First, clinical (and food) microbiology laboratories could perform reflex culturing of biological specimens (or food) that test positive via CIDTs such that positive isolates can still be submitted to public health laboratories for subtyping or other culture-based tests such as WGS (Cronquist et al., 2012). Second, clinical laboratories could alternatively submit biological specimens to public health laboratories that have tested positive using CIDTs. Lastly, culture-independent subtyping and streamlined bioinformatics analyses could be developed for both public health and clinical

microbiology laboratories; this would be particularly beneficial for specimens incompatible or not optimal with culture (e.g., fecal swabs; Kotton et al., 2006). Hence, while any of these methods would require substantial restructuring of national and international surveillance infrastructure it would in theory overcome the prospective predicament whereby public health laboratories would not have access to a considerable portion of isolates and thus capacity for precise outbreak detection and source attribution. If, however, culture were rendered obsolete prior to the implementation of any of the above-listed scenarios, it would inevitably be detrimental for public health surveillance. Additionally, the absence of isolates (and a biobank of historical isolates) will make retrospective studies of outbreaks challenging. In this regard, regulatory groups have been created in both Canada and the US to address the concern pertaining to a lack of enteric isolates available for further characterizations at public health laboratories. The US has created an interim recommendations document⁸ to ensure that isolation is attempted or that positive CIDT specimens are retained – Canada is in the process of developing analogous guidelines.

TRANSFORMATION OF ANTIMICROBIAL SUSCEPTIBILITY TESTING AND INFLUENCE ON ANTIMICROBIAL RESISTANCE SURVEILLANCE SYSTEMS

A limited number of conventional growth-based methods for AST have persisted in routine usage throughout the transformation of diagnostic microbiology. Included among these are disk diffusion (e.g., Kirby-Bauer) strategies and broth microdilutions. The latter of which has achieved gold standard status; thus, novel AST methods are compared to the efficacy of broth microdilutions from development through clinical trial. At present, AST is either accomplished via the above-listed conventional manual methods or growth-dependent automated AST systems including the Vitek System (bioMerieux, France), Avantage Test System (Abbott Laboratories, Irving, TX, United States), Phoenix (BD Biosciences, Cockeysville, MD, United States) and others, all of which are based on broth microdilution testing (Van Belkum and Dunne, 2013). More recent growth-based AST techniques have also been developed that generally employ innovative methods; these include MALDI-TOF mass spectrometry, microfluidics (NanoDrop BMD), isothermal microcalorimetry, real-time microscopy, and others.

Whole Genome Sequencing Antimicrobial Resistance Gene Detection

With the rising usage of molecular CIDTs and NGS strategies in clinical diagnostics, speculation exists regarding the potential feasibility of NGS methods or other advanced technologies, as aforementioned, replacing growth-based AST. Assessments have

⁸https://www.aphl.org/aboutAPHL/publications/Documents/FS-Enteric_Pathogens_Guidelines_0216.pdf

been completed for some species [e.g., *Salmonella* (McDermott et al., 2016) and *Gonorrhea* (Demczuk et al., 2015)] and it has been suggested that resistance genes can be accurately detected and further, the presence of genes and mutations associated with AMR has been shown to have a high correlation with phenotypic antimicrobial susceptibility profiles. As an example, WGS was recently applied to non-typhoidal *Salmonella* isolates (McDermott et al., 2016). The authors reported an overall 99% concordance rate between genotype and phenotype; for most classes of antibiotics, concordance was closer to 100% while lower for aminoglycosides and beta-lactams. Though inferring AST via WGS is becoming more common in diagnostic laboratories, to date, few studies have performed large-scale sequencing projects to investigate the utility of WGS to complement or replace conventional AST in routine laboratory workflows. Nonetheless, limitations concerning particular phenotypic-only traits that can't be determined exclusively from WGS or other NGS methodologies will need to be addressed.

Metagenomics Antimicrobial Resistance Gene Detection

As discussed previously, conventional approaches used to determine AMR are generally reliant upon growth-based tests and further, are heavily targeted to testing human pathogens. The methodological spectrum for AMR gene detection must be expanded to overcome imperative limitations: relatively few bacterial microbes can be cultured (Eckburg et al., 2005) and commensals are thought to comprise a resistance gene pool (resistome) that can be transferred to pathogens (Davies and Davies, 2010). Hence, metagenomics has the capacity to overcome challenges associated with traditional AST. At present, both shotgun metagenomics sequencing and functional metagenomics have been utilized to interrogate the resistome.

The use of shotgun metagenomics sequencing to identify AMR genes is promising and studies are now beginning to elucidate their plausible role in surveillance. In this technique, metagenomics reads are mapped against a database containing a comprehensive catalog of known AMR genes; notable examples include CARD (Jia et al., 2017), ARDB (Liu and Pop, 2009), Resfams (Gibson et al., 2014), and MEGARes (Lakin et al., 2017). Alternatively, metagenomics reads can be assembled into contigs and next compared to a functional annotation database. A recent study surveyed the metagenomes of several ecological niches including the human gastrointestinal tract, water, animals and others for AMR genes (Fitzpatrick et al., 2016). A large abundance of AMR genes were detected in the human gastrointestinal tract and variably identified in other metagenomes. Nonetheless, the authors emphasized the importance of difficulties associated with shotgun metagenomics detection of AMR genes. First, similar to determining microbial abundance reference ranges (from a diagnostics perspective), limits of detection need to be established in order to have sufficient coverage with the capacity to detect rare AMR genes or in complex metagenomes. Second, methods to normalize data to overcome variable microbial diversity and genome sizes will also be required. Third, the biological features inherent of AMR genes

(e.g., often carried within the mobilome or other transmissible genetic elements) render sequencing and data analysis difficult due to their repetitive nature.

Numerous studies have applied shotgun metagenomics sequencing to detect AMR genes. The *C. difficile* study discussed above identified a total of 27 AMR genes and 55.6% of samples contained a minimum of 1 gene (Zhou et al., 2016). The most dominant AMR genes encoded cephalosporin (*Bl2e_cfxa*; 25.9%) and tetracycline (*tetQ*; 25.9%) resistance whereas macrolide (*ermA*, *ermB*, *ermF*, *ermG*) resistance was variable ranging from 3.7 to 11.1% of samples. It was also shown that WGS of isolates predicted AMR phenotypes with high accuracy. A recent study explored the utility of shotgun metagenomics to detect MDR pediatric bacterial infections, specifically, methicillin-resistant *Staphylococcus aureus*, vancomycin-resistant *Enterococcus* and MDR Enterobacteriaceae (Andersen et al., 2016). The study included three cohorts – high-risk inpatients, low-risk outpatients and controls. Though the potential for MDR bacteria was increased in inpatients and outpatients compared to controls, no differences were detected between inpatients and outpatients. A noteworthy observation was that 53% of inpatients were colonized with an MDR bacterium that culture failed to identify.

Functional metagenomics on the other hand is an exceedingly powerful technique attributed to its capacity to discover novel and highly divergent AMR genes (Perry and Wright, 2014). This method involves cloning total community genomic DNA into an expression vector and transformation into a susceptible expression host (e.g., *E. coli*). The transformant library is then assayed for AMR by culturing on selective media – persistent AMR genes are sequenced and annotated. Therefore, this method is advantageous, ascribed to the possibility to determine genotypic and phenotypic traits.

As NGS-based methodologies like WGS are proving useful in clinical and public health laboratories, we expect that other advanced sequencing techniques will similarly be effective in amalgamating molecular AST and phenotypic susceptibilities thereby allowing for a complete bypass of microbial culture. In this regard, NGS supplemented with metatranscriptomics or proteomics can be more informative as it allows for description of both gene presence and expression (Cohen et al., 2015; Perez-Llarena and Bou, 2016). A combinatorial ‘omics’ approach incorporating any of the above-mentioned assays may be initially computationally laborious due to a paucity of streamlined analysis techniques and standardization. However, upon analytical validation, such an approach may still be less laborious than conventional culture-dependent testing.

Current Condition of Antimicrobial Resistance Surveillance

Public health laboratories routinely track specific characteristics of bacterial pathogens that are implicated in infection, thus effectively allowing for increased understanding of bacterial pathogens and their epidemiology. Among these characteristics include monitoring AMR of common enteric pathogens such as *Salmonella* spp. and *Campylobacter* spp., as well as virulence profiles in for example STEC. At the Canadian public health

level, AMR surveillance is conducted through the Canadian Integrated Program for Antimicrobial Resistance Surveillance (CIPARS)⁹ in combination with FoodNet. Importantly, though AMR in a clinical setting is largely associated with the utilization of antimicrobials for the treatment of infections, their use in agri-food production is also known to contribute to the resistant microbe pool. Resistant bacteria are associated with more severe disease (van Duin and Paterson, 2016) and poor patient outcome, hence monitoring for AMR is critical. Accordingly, AMR or MDR bacteria have caused several recent foodborne outbreaks (Cartwright et al., 2016; Gieraltowski et al., 2016; Kawakami et al., 2016).

Irrespective of how and when NGS technologies will be capable of rapidly and accurately detecting AMR and associated susceptibilities, difficulties in public health AMR surveillance will be apparent. With a growing number of clinical and food microbiology laboratories opting for utilization of CIDTs for diagnostic purposes, laboratories are bypassing the need to culture microbes; this may affect public health surveillance of AMR trends which continue to use isolates as the foundation for AMR surveillance programs and assessment of phenotypic resistance to antimicrobials. CIDT for enteric pathogens that yield no isolate and potentially no specimen for public health laboratories does pose an important challenge, however, actions can be taken to maintain adequate AMR surveillance. Gonorrhea, for example, represents a proof-of-principle for overcoming a lack of isolates in the context of surveillance. In particular, routine testing of gonorrhea has been performed by NAATs for some time due to the higher sensitivity and specificity. Guidelines have been documented regarding sentinel surveillance mechanisms to continue monitoring AMR trends¹⁰. Moreover, CIDT techniques could also be developed that have the capacity to test for known AMR though culture will still be systematically needed to detect novel and emerging resistance mechanisms.

FUTURE CHALLENGES OF METAGENOMICS IN DIAGNOSTIC AND PUBLIC HEALTH SETTINGS

Whole genome sequencing as a complement to conventional culture-based, molecular or serological CIDTs in real-time serves as an empirical model for the use of future techniques (Jackson et al., 2016). From lessons learnt via WGS application, metagenomics and other NGS techniques should move into front-line laboratories in the near future. Initial proof-of-concept studies (Gilmour et al., 2010), retrospective analyses (Loman et al., 2013), validation processes and real-time implementation of WGS (Jackson et al., 2016) in public health surveillance and outbreak response will undoubtedly help enhance acceptance of the aforementioned metagenomics methodologies from the medical and public health community at large. Standardization is an essential element for implementation of any method set to become the next “gold standard.” Various WGS studies have

been undertaken of historical clusters and sporadic cases to determine concordance of WGS data with traditional subtyping and epidemiological data; in many instances, WGS provided higher discriminatory power than traditional subtyping methods and thus resulted in many occurrences whereby isolates were either included or excluded from an outbreak due to WGS and supportive epidemiological data, if present. To date, isolate cluster inclusion/exclusion criteria remains a moving target due to sizeable differences in biology amongst foodborne pathogens. Although WGS is part of the public health toolbox, it remains to be standardized in such a way that it can be applied to all foodborne pathogens. As a result, regular cluster audits to assess analysis pipelines for performance in addition to adequacy of quality control and assurance metrics for sequencing and all analysis outputs should be routinely conducted. Furthermore, the transition to add WGS as a public health epidemiology tool has been an internationally driven effort with organizations and partners from every field – primary health care, academia, industry and all government tiers. For wide adoption of new methods, involvement of the international community is key. Each of these aforementioned guidelines will be paramount to ensure cluster definitions remain robust in the context of genomic epidemiology. Another crucial aspect of implementing new technologies is knowledge translation. Currently, large efforts are being directed at educating and communicating the use of WGS in public health surveillance and outbreak response. Not only will laboratory technicians be required to advance their skillsets, but also policymakers and media will require training on how to interpret highly technical data and communicate it to the public (Jackson et al., 2016). As we enter this new era of “omics” in public health, the aforementioned criteria will be instrumental in guiding and implementing unbiased NGS-based CIDT methods in foodborne disease surveillance and outbreak response.

As mentioned previously, preceding widespread implementation of metagenomics sequencing and analysis in diagnostic and public health settings, bioinformatics encompassing algorithms, software and pipelines will need to be developed, validated and standardized through various ring trials and retrospective analyses. Perhaps the largest obstacle faced by the field of metagenomics is the lack of universal analyses or pipeline recognized as the status quo. In this regard, we previously discussed the significance of curated sequence databases to appropriate patient diagnosis. The GMI¹¹ consortium has begun to address this issue via the creation of a global database that house uniquely identifiable microbial genomic data in combination with high quality meta- and epidemiological data. This drive will assist in the global amalgamation of microbial WGS and metadata in an easily accessible portal ultimately enhancing global surveillance of infectious microbes and emergent pathogens. Moreover, such a global database will similarly be influential in data mining investigations as they relate to gene level comparisons such as AMR, virulence, and environmental fitness.

Procurement of regulatory approval for routine implementation of NGS-based technologies in diagnostic, public

⁹<http://www.phac-aspc.gc.ca/cipars-picra/index-eng.php>

¹⁰<http://www.phac-aspc.gc.ca/std-mts/sti-its/cgsti-ldcits/>

¹¹<https://www.globalmicrobialidentifier.org>

health and food safety laboratories will be onerous owing to various difficulties inherent of such an assay. First, metagenomics dataset contain identifiable information. As discussed previously, many biological specimen types have high amounts of host DNA; thus, a large proportion of generated sequences will be of eukaryotic origin. In research- and diagnostic-based cases, host DNA is filtered from the dataset, however, issues are apparent with how best to protect patient privacy. There may be plausibility that genetically informative host sequence data could be used to screen against a panel of known disease-causing genetic variants. Providing patients with information pertaining to a potential genetic disease via such an assay is an ethical concern (Chrystoja and Diamandis, 2014). On the other hand, providing patients with their metagenomes may lead to precarious implications should patients seek to self-diagnose. Second, access to patent DNA from the human genome, food products and ingredients sequence data could pose unwarranted legal implications. The recovery of patented DNA sequences from food sources or ingredients such as genetically modified foods (crops or animals) may give rise to concerns of patent infringement (Chrystoja and Diamandis, 2014). Metagenomics may also be used for the screening detection of unauthorized GMO use and similarly aid regulation of the food industry, particularly with high import/export rate of global food products and ingredients. Additionally, metagenomics may allow for the detection of species fraud, food product mislabelling or incorrect claims (Corrado, 2016) that may be subject to increased scrutiny. Third, the detection of AMR or virulence genes may lead to undesirable outcomes. Specifically, rising concerns with AMR has led to the complete ban of antimicrobial sub-therapeutic use in food animals in the European Union. The potential to identify AMR in food products may inhibit producers and commercial food processors from agreeing to regulations or policies related to metagenomics testing of food products in fear of repercussions in food trade and export. Overall, should metagenomics become a validated assay in clinical, food and public health settings its impact may be far-reaching in the realm of ethical and legal implications.

REFERENCES

- Ahn, T. H., Chai, J., and Pan, C. (2015). Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* 31, 170–177. doi: 10.1093/bioinformatics/btu641
- Alfa, M. J., Strang, D., Tappia, P. S., Graham, M., Van Domselaar, G., Forbes, J. D., et al. (2017). A randomized trial to determine the impact of a digestion resistant starch composition on the gut microbiome in older and mid-age adults. *Clin. Nutr.* doi: 10.1016/j.clnu.2017.03.025 [Epub ahead of print].
- Allard, M. W., Strain, E., Melka, D., Bunning, K., Musser, S. M., Brown, E. W., et al. (2016). Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J. Clin. Microbiol.* 54, 1975–1983. doi: 10.1128/JCM.00081-16
- Almonacid, D. E., Kraal, L., Ossandon, F. J., Budovskaya, Y. V., Cardenas, J. P., Richman, J., et al. (2016). 16S rRNA gene sequencing as a clinical diagnostic aid for gastrointestinal-related conditions. *bioRxiv*. doi: 10.1101/084657
- Andersen, H., Connolly, N., Bangar, H., Staat, M., Mortensen, J., Deburger, B., et al. (2016). Use of shotgun metagenome sequencing to detect fecal colonization with multidrug-resistant bacteria in children. *J. Clin. Microbiol.* 54, 1804–1813. doi: 10.1128/JCM.02638-15
- Anthony, R. M., Brown, T. J., and French, G. L. (2000). Rapid diagnosis of bacteremia by universal amplification of 23S ribosomal DNA followed by hybridization to an oligonucleotide array. *J. Clin. Microbiol.* 38, 781–788.
- Anthony, S. J., Epstein, J. H., Murray, K. A., Navarrete-Macias, I., Zambrana-Torrello, C. M., Solovyov, A., et al. (2013). A strategy to estimate unknown viral diversity in mammals. *mBio* 4:e00598–13. doi: 10.1128/mBio.00598-13
- Argenio, V. D., Casaburi, G., Colicchio, R., Sarnataro, D., Discepolo, V., Kim, S. M., et al. (2016). Mucosal gut microbiome is associated with celiac disease-specific microbiome alteration in adult patients. *Am. J. Gastroenterol.* 111, 1659–1661. doi: 10.1038/ajg.2016.227
- Ashton, P. M., Nair, S., Peters, T. M., Bale, J. A., Powell, D. G., Painset, A., et al. (2016). Identification of *Salmonella* for public health surveillance using whole genome sequencing. *PeerJ* 4:e1752. doi: 10.7717/peerj.1752
- Aw, T. G., Wengert, S., and Rose, J. B. (2016). Metagenomic analysis of viruses associated with field-grown and retail lettuce identifies human and animal viruses. *Int. J. Food Microbiol.* 223, 50–56. doi: 10.1016/j.ijfoodmicro.2016.02.008
- Barratt, J. L. N., Harkness, J., Marriott, D., Ellis, J. T., and Stark, D. (2011). A review of *Dientamoeba fragilis* carriage in humans: several reasons why this organism

CONCLUSION

The prospective use of diagnostic metagenomics and comparable techniques offers an assumption-free workflow (though biases are present with each methodology; **Table 1**) thus creating the ability to detect any and all pathogens (bacteria, virus, parasite, and others) from various biological specimens or food products. Through forthcoming improvements related to required technical expertise, throughput and cost-effectiveness of sequencing combined with enhanced and streamlined laboratory and bioinformatics pipelines, metagenomics will likely have a predominant role in the diagnostic and public health laboratory. We expect an automated metagenomics pipeline will complement and may even replace several methods currently employed in the diagnostic laboratory while concurrently providing additional information such as AMR, virulence, and genomic epidemiology. It is apparent that the functionality of diagnostic metagenomics has been established in research settings and further, in detecting etiological culprits of unidentified illnesses and outbreaks. We anticipate that within the next decade, detection and characterization of pathogens via metagenomics-based workflows will be implemented in routine usage in diagnostic and public health laboratories.

AUTHOR CONTRIBUTIONS

NK and AR developed the concept for the manuscript. JF, NK, JR, FP, and AR wrote the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Celine Nadon for her expert review of the manuscript. A figure in this paper used icons made by Freepik, from www.flaticon.com.

- should be considered in the diagnosis of gastrointestinal illness. *Gut Microbes* 2, 3–12. doi: 10.4161/gmic.2.1.14755
- Bavelaar, H. H., Rahamat-Langendoen, J., Niesters, H. G., Zoll, J., and Melchers, W. J. (2015). Whole genome sequencing of fecal samples as a tool for the diagnosis and genetic characterization of norovirus. *J. Clin. Virol.* 72, 122–125. doi: 10.1016/j.jcv.2015.10.003
- Bergholz, T. M., Moreno Switt, A. I., and Wiedmann, M. (2014). Omics approaches in food safety: fulfilling the promise? *Trends Microbiol.* 22, 275–281. doi: 10.1016/j.tim.2014.01.006
- Brown, J. R., Roy, S., Ruis, C., Yara Romero, E., Shah, D., and Williams, R. (2016). Norovirus whole-genome sequencing by SureSelect target enrichment: a robust and sensitive method. *J. Clin. Microbiol.* 54, 2530–2537. doi: 10.1128/JCM.01052-16
- Buchan, B. W., Olson, W. J., Pezewski, M., Marcon, M. J., Novicki, T., Uphoff, T. S., et al. (2013). Clinical evaluation of a real-time PCR assay for identification of *Salmonella*, *Shigella*, *Campylobacter* (*Campylobacter jejuni* and *C. coli*), and shiga toxin-producing *Escherichia coli* isolates in stool specimens. *J. Clin. Microbiol.* 51, 4001–4007. doi: 10.1128/JCM.02056-13
- Buss, S. N., Alter, R., Iwen, P. C., and Fey, P. D. (2013). Implications of culture-independent panel-based detection of *Cyclospora cayetanensis*. *J. Clin. Microbiol.* 51, 3909. doi: 10.1128/JCM.02238-13
- Buss, S. N., Leber, A., Chapin, K., Fey, P. D., Bankowski, M. J., Jones, M. K., et al. (2015). Multicenter evaluation of the BioFire FilmArray gastrointestinal panel for etiologic diagnosis of infectious gastroenteritis. *J. Clin. Microbiol.* 53, 915–925. doi: 10.1128/JCM.02674-14
- Carleton, H. A., and Gerner-Smidt, P. (2016). Whole-genome sequencing is taking over foodborne disease surveillance. *Microbe* 11, 311–317.
- Cartwright, E. J., Nguyen, T., Melluso, C., Ayers, T., Lane, C., Hodges, A., et al. (2016). A multistate investigation of antibiotic-resistant *Salmonella enterica* serotype I 4,[5], 12:i:- infections as part of an international outbreak associated with frozen feeder rodents. *Zoonoses Public Health* 63, 62–71. doi: 10.1111/zph.12205
- Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., and Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* 73, 278–288. doi: 10.1128/AEM.01177-06
- Chattaway, M. A., Dallman, T. J., Gentle, A., Wright, M. J., Long, S. E., Ashton, P. M., et al. (2016). Whole genome sequencing for public health surveillance of shiga toxin-producing *Escherichia coli* other than serogroup O157. *Front. Microbiol.* 7:258. doi: 10.3389/fmicb.2016.00258
- Chrystoja, C. C., and Diamandis, E. P. (2014). Whole genome sequencing as a diagnostic test: challenges and opportunities. *Clin. Chem.* 60, 724–733. doi: 10.1373/clinchem.2013.209213
- Cinar, H. N., Gopinath, G., Jarvis, K., and Murphy, H. R. (2015). The complete mitochondrial genome of the foodborne parasitic pathogen *Cyclospora cayetanensis*. *PLoS ONE* 10:e0128645. doi: 10.1371/journal.pone.0128645
- Cleary, B., Brito, I. L., Huang, K., Gevers, D., Shea, T., Young, S., et al. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* 33, 1053–1060. doi: 10.1038/nbt.3329
- Cohen, A., Bont, L., Engelhard, D., Moore, E., Fernández, D., Kreisberg-Greenblatt, R., et al. (2015). A multifaceted “omics” approach for addressing the challenge of antimicrobial resistance. *Fut. Microbiol.* 10, 365–376. doi: 10.2217/fmb.14.127
- Corrado, G. (2016). Advances in DNA typing in the agro-food supply chain. *Trends Food Sci. Technol.* 52, 80–89. doi: 10.1016/j.tifs.2016.04.003
- Cronquist, A. B., Mody, R. K., Atkinson, R., Besser, J., D’Angelo, M. T., Hurd, S., et al. (2012). Impacts of culture-independent diagnostic practices on public health surveillance for bacterial enteric pathogens. *Clin. Infect. Dis.* 54, S432–S439. doi: 10.1093/cid/cis267
- Davies, J., and Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* 74, 417–433. doi: 10.1128/mmbr.00016-10
- Demczuk, W., Lynch, T., Martin, I., Van Domselaar, G., Graham, M., Bharat, A., et al. (2015). Whole-genome phylogenomic heterogeneity of *Neisseria gonorrhoeae* isolates with decreased cephalosporin susceptibility collected in Canada between 1989 and 2013. *J. Clin. Microbiol.* 53, 191–200. doi: 10.1128/JCM.02589-14
- Denno, D. M., Shaikh, N., Stapp, J. R., Qin, X., Hutter, C. M., Hoffman, V., et al. (2012). Diarrhea etiology in a pediatric emergency department: a case control study. *Clin. Infect. Dis.* 55, 897–904. doi: 10.1093/cid/cis553
- Doggett, N. A., Mukundan, H., Lefkowitz, E. J., Slezak, T. R., Chain, P. S., Morse, S., et al. (2016). Culture-independent diagnostics for health security. *Health Secur.* 14, 122–142. doi: 10.1089/hs.2015.0074
- Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., et al. (2005). Diversity of the human intestinal microbial flora. *Science* 308, 1635–1638. doi: 10.1126/science.1110591
- Escobar-Zepeda, A., Sanchez-Flores, A., and Quirasco Baruch, M. (2016). Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. *Food Microbiol.* 57, 116–127. doi: 10.1016/j.fm.2016.02.004
- Ettayebi, K., Crawford, S. E., Murakami, K., Broughman, J. R., Karandikar, U., Tenge, V. R., et al. (2016). Replication of human noroviruses in stem cell-derived human enteroids. *Science* 353, 1387–1393. doi: 10.1126/science.aaf5211
- Finkbeiner, S. R., Allred, A. F., Tarr, P. I., Klein, E. J., Kirkwood, C. D., and Wang, D. (2008). Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog.* 4:e1000011. doi: 10.1371/journal.ppat.1000011
- Finkbeiner, S. R., Li, Y., Ruone, S., Conrardy, C., Gregoricus, N., Toney, D., et al. (2009). Identification of a novel astrovirus (astrovirus VA1) associated with an outbreak of acute gastroenteritis. *J. Virol.* 83, 10836–10839. doi: 10.1128/JVI.00998-09
- Fitzgerald, C., Patrick, M., Gonzalez, A., Akin, J., Polage, C. R., Wymore, K., et al. (2016). Multicenter evaluation of clinical diagnostic methods for detection and isolation of *Campylobacter* spp. from stool. *J. Clin. Microbiol.* 54, 1209–1215. doi: 10.1128/JCM.01925-15.Editor
- Fitzpatrick, D., Walsh, F., and Walsh, F. (2016). Antibiotic resistance genes across a wide variety of metagenomes. *FEMS Microbiol. Ecol.* 92:fiv168. doi: 10.1093/femsec/fiv168
- Flygare, S., Simon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., et al. (2016). Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol.* 17, 111. doi: 10.1186/s13059-016-0969-1
- Forbes, J. D., Van Domselaar, G., and Bernstein, C. N. (2016). Microbiome survey of the inflamed and noninflamed gut at different compartments within the gastrointestinal tract of inflammatory bowel disease patients. *Inflamm. Bowel Dis.* 22, 817–825. doi: 10.1097/MIB.0000000000000684
- Forster, S. C., Anonye, B. O., Kumar, N., Neville, B. A., Stares, M. D., Goulding, D., et al. (2016). Culturing of “unculturable” human microbiota reveals novel taxa and extensive sporulation. *Nature* 533, 543–546. doi: 10.1038/nature17645
- Gambarin, P., Magnabosco, C., Losio, M. N., Pavoni, E., Gattuso, A., Arcangeli, G., et al. (2012). *Listeria monocytogenes* in ready-to-eat seafood and potential hazards for the consumers. *Int. J. Microbiol.* 2012:497635. doi: 10.1155/2012/497635
- Gao, D., Yu, Q., Wang, G., Wang, G., and Xiong, F. (2016). Diagnosis of a malayan filariasis case using a shotgun diagnostic metagenomics assay. *Parasit. Vectors* 9, 86. doi: 10.1186/s13071-016-1363-2
- Gentry-Shields, J., and Jaykus, L.-A. (2015). Comparison of process control viruses for use in extraction and detection of human norovirus from food matrices. *Food Res. Int.* 77, 320–325. doi: 10.1016/j.foodres.2015.05.027
- Gibson, M. K., Forsberg, K. J., and Dantas, G. (2014). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* 9, 207–216. doi: 10.1038/ismej.2014.106
- Gieraltowski, L., Higa, J., Peralta, V., Green, A., Schwensohn, C., Rosen, H., et al. (2016). National outbreak of multidrug resistant *Salmonella* heidelberg infections linked to a single poultry company. *PLoS ONE* 11:e0162369. doi: 10.1371/journal.pone.0162369
- Gill, A. (2017). The importance of bacterial culture to food microbiology in the age of genomics. *Front. Microbiol.* 8:777. doi: 10.3389/fmicb.2017.00777
- Gilmour, M. W., Graham, M., Van Domselaar, G., Tyler, S., Kent, H., Trout-yakel, K. M., et al. (2010). High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. *BMC Genomics* 11:120. doi: 10.1186/1471-2164-11-120
- Gorski, L. (2012). Selective enrichment media bias the types of *Salmonella enterica* strains isolated from mixed strain cultures and complex enrichment broths. *PLoS ONE* 7:e34722. doi: 10.1371/journal.pone.0034722
- Hahn, B., Maldonado, Y., and Schreiber, E. (2003). Subtyping of foodborne and environmental isolates of *Escherichia coli* by multiplex-PCR, rep-PCR, PFGE,

- ribotyping and AFLP. *J. Microbiol. Methods* 53, 387–399. doi: 10.1016/S0167-7012(02)00259-2
- Hanson, K. E., and Couturier, M. R. (2016). Multiplexed molecular diagnostics for respiratory, gastrointestinal, and central nervous system infections. *Clin. Infect. Dis.* 63, 1361–1367. doi: 10.1093/cid/ciw494
- Harrington, S. M., Buchan, B. W., Doern, C., Fader, R., Ferraro, M. J., Pillai, D. R., et al. (2015). Multicenter evaluation of the BD max enteric bacterial panel PCR assay for rapid detection of *Salmonella* spp., *Shigella* spp., *Campylobacter* spp. (*C. jejuni* and *C. coli*), and shiga toxin 1 and 2 genes. *J. Clin. Microbiol.* 53, 1639–1647. doi: 10.1128/JCM.03480-14
- Hasan, M. R., Rawat, A., Tang, P., Jithesh, P. V., Thomas, E., Tan, R., et al. (2016). Depletion of human DNA in spiked clinical specimens to improve the sensitivity of pathogen detection by next generation sequencing. *J. Clin. Microbiol.* 54, 919–927. doi: 10.1128/JCM.03050-15
- Hasman, H., Saputra, D., Sicheritz-Ponten, T., Lund, O., Svendsen, C. A., Frimodt-Møller, N., et al. (2014). Rapid whole genome sequencing for the detection and characterization of microorganisms directly from clinical samples. *J. Clin. Microbiol.* 52, 139–146. doi: 10.1128/JCM.02452-13
- Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., et al. (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* 2, 33. doi: 10.1186/2049-2618-2-33
- Huang, J. Y., Henao, O. L., Griffin, P. M., Vugia, D. J., Cronquist, A. B., Hurd, S., et al. (2016). Infection with pathogens transmitted commonly through food and the effect of increasing use of culture-independent diagnostic tests on surveillance - foodborne diseases active surveillance network, 10 U.S. sites, 2012–2015. *Morb. Mortal. Wkly. Rep.* 65, 368–371. doi: 10.15585/mmwr.mm6514a2
- Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234
- Jackson, B. R., Tarr, C., Strain, E., Jackson, K. A., Conrad, A., Carleton, H., et al. (2016). Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. *Clin. Infect. Dis.* 63, 380–386. doi: 10.1093/cid/ciw242
- Jandaa, J. M., and Abbott, S. A. (2014). Culture-independent diagnostic testing: have we opened Pandora's box for good? *Diagn. Microbiol. Infect. Dis.* 80, 171–176. doi: 10.1016/j.diagmicrobio.2014.08.001
- Jarvis, K. G., White, J. R., Grim, C. J., Ewing, L., Ottesen, A. R., Beaubrun, J. J.-G., et al. (2015). Cilantro microbiome before and after nonselective pre-enrichment for *Salmonella* using 16S rRNA and metagenomic sequencing. *BMC Microbiol.* 15:160. doi: 10.1186/s12866-015-0497-2
- Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., et al. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 45, D566–D573. doi: 10.1093/nar/gkw1004
- Jones, M. K., Watanabe, M., Zhu, S., Graves, C. L., Keyes, L. R., Grau, K. R., et al. (2014). Enteric bacteria promote human and mouse norovirus infection of B cells. *Science* 346, 755–759. doi: 10.1126/science.1257147
- Joseph, S. J., Li, B., Petit, R. A. III, Qin, Z. S., Darrow, L., and Read, T. D. (2016). The single-species metagenome: subtyping *Staphylococcus aureus* core genome sequences from shotgun metagenomic data. *PeerJ* 4, e2571. doi: 10.7717/peerj.2571
- Jung, J. Y., Lee, S. H., Kim, J. M., Park, M. S., Bae, J., Hahn, Y., et al. (2011). Metagenomic analysis of kimchi, a traditional korean fermented food. *Appl. Environ. Microbiol.* 77, 2264–2274. doi: 10.1128/AEM.02157-10
- Kataoka, T., Yamaguchi, H., Sato, M., Watanabe, T., Taniuchi, Y., Kuwata, A., et al. (2016). Seasonal and geographical distribution of near-surface small photosynthetic-eukaryotes in the western North Pacific determined by pyrosequencing of 18S rDNA. *FEMS Microbiol. Ecol.* 93, fiw229. doi: 10.1093/femsec/fiw229
- Kawai, T., Sekizuka, T., Yahata, Y., Kuroda, M., Kumeda, Y., Iijima, Y., et al. (2012). Identification of *Kudoa septempunctata* as the causative agent of novel food poisoning outbreaks in Japan by consumption of *Paralichthys olivaceus* in raw fish. *Clin. Infect. Dis.* 54, 1046–1052. doi: 10.1093/cid/cir1040
- Kawakami, V. M., Botticchio, L., Angelo, K., Linton, N., Kissler, B., Basler, C., et al. (2016). Notes from the field: outbreak of multidrug-resistant *Salmonella* infections linked to pork – Washington, 2015. *Morb. Mortal. Wkly. Rep.* 65, 379–381. doi: 10.15585/mmwr.mm6514a4
- Kleiner, M., Hooper, L. V., and Duerkop, B. A. (2015). Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* 16:7. doi: 10.1186/s12864-014-1207-4
- Korsak, N., Taminiau, B., Leclercq, M., Nezer, C., Crevecoeur, S., Ferauche, C., et al. (2015). Short communication: evaluation of the microbiota of kefir samples using metagenetic analysis targeting the 16S and 26S ribosomal DNA fragments. *J. Dairy Sci.* 98, 3684–3689. doi: 10.3168/jds.2014-9065
- Köser, C. U., Ellington, M. J., Cartwright, E. J. P., Gillespie, S. H., Brown, N. M., Farrington, M., et al. (2012). Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog.* 8:e1002824. doi: 10.1371/journal.ppat.1002824
- Kotton, C. N., Lankowski, A. J., and Hohmann, E. L. (2006). Comparison of rectal swabs with fecal cultures for detection of *Salmonella typhimurium* in adult volunteers. *Diagn. Microbiol. Infect. Dis.* 56, 123–126. doi: 10.1016/j.diagmicrobio.2006.04.003
- Kundu, S., Lockwood, J., Depledge, D. P., Chaudhry, Y., Aston, A., Rao, K., et al. (2013). Next-generation whole genome sequencing identifies the direction of norovirus transmission in linked patients. *Clin. Infect. Dis.* 57, 407–414. doi: 10.1093/cid/cit287
- Lakin, S. M., Dean, C., Noyes, N. R., Dettenwanger, A., Ross, A. S., Doster, E., et al. (2017). MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res.* 45, D574–D580. doi: 10.1093/nar/gkw1009
- Lambert, D., Carrillo, C. D., Koziol, A. G., Manninger, P., and Blais, B. W. (2015). GeneSippr: a rapid whole-genome approach for the identification and characterization of foodborne pathogens such as priority Shiga toxicigenic *Escherichia coli*. *PLoS ONE* 10:e0122928. doi: 10.1371/journal.pone.0122928
- Langelier, C., Zinter, M., Kalantar, K., Yanik, G., Christenson, S., Odonovan, B., et al. (2017). Metagenomic next-generation sequencing detects pulmonary pathogens in hematopoietic cellular transplant patients with acute respiratory illnesses. *bioRxiv*. doi: 10.1101/102798
- Law, J. W., Mutualib, N. A., Chan, K., Lee, L., and Lee, L. (2015). An insight into the isolation, enumeration, and molecular detection of *Listeria monocytogenes* in food. *Front. Microbiol.* 6:1227. doi: 10.3389/fmicb.2015.01227
- Law, J. W. F., Mutualib, N. S., Chan, K. G., and Lee, L. H. (2014). Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. *Front. Microbiol.* 5:770. doi: 10.3389/fmicb.2014.00770
- Leimena, M. M., Ramiro-garcia, J., Davids, M., Bogert, B., Van Den Smidt, H., Smid, E. J., et al. (2013). A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC Genomics* 14:530. doi: 10.1186/1471-2164-14-530
- Leite, A. M. O., Mayo, B., Rachid, C. T. C. C., Peixoto, R. S., Silva, J. T., Paschoalin, V. M. F., et al. (2012). Assessment of the microbial diversity of Brazilian kefir grains by PCR-DGGE and pyrosequencing analysis. *Food Microbiol.* 31, 215–221. doi: 10.1016/j.fm.2012.03.011
- Leonard, S. R., Mammel, M. K., Lacher, D. W., and Elkins, C. A. (2015). Application of metagenomic sequencing to food safety: detection of shiga toxin-producing *Escherichia coli* on fresh bagged spinach. *Appl. Environ. Microbiol.* 81, 8183–8191. doi: 10.1128/AEM.02601-15
- Lesho, E., Clifford, R., Onmus-Leone, F., Appalla, L., Snetsrud, E., Kwak, Y., et al. (2016). The challenges of implementing next generation sequencing across a large healthcare system, and the molecular epidemiology and antibiotic susceptibilities of carbapenemase-producing bacteria in the healthcare system of the U.S. Department of Defense. *PLoS ONE* 11:e0155770. doi: 10.1371/journal.pone.0155770
- Lienau, E. K., Strain, E., Wang, C., Zheng, J., Ottesen, A. R., Keys, C. E., et al. (2011). Identification of a salmonellosis outbreak by means of molecular sequencing. *N. Engl. J. Med.* 364, 981–982. doi: 10.1056/NEJMci1100443
- Liu, B., and Pop, M. (2009). ARDB - Antibiotic resistance genes database. *Nucleic Acids Res.* 37, 443–447. doi: 10.1093/nar/gkn656
- Loman, N. J., Constantinidou, C., Christner, M., Rohde, H., Chan, J. Z.-M., Quick, J., et al. (2013). A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic *Escherichia coli* O104:H4. *J. Am. Med. Assoc.* 309, 1502–1510. doi: 10.1001/jama.2013.3231
- Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nat. Protoc.* 12, 213–218. doi: 10.1038/nprot.2016.182

- Margot, H., Stephan, R., and Tasara, T. (2016). Mungo bean sprout microbiome and changes associated with culture based enrichment protocols used in detection of Gram-negative foodborne pathogens. *Microbiome* 4, 48. doi: 10.1186/s40168-016-0193-y
- Mason, O. U., Scott, N. M., Gonzalez, A., Robbins-Pianka, A., Bælum, J., Kimbrel, J., et al. (2014). Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *ISME J.* 8, 1464–1475. doi: 10.1038/ismej.2013.254
- McCollum, J. T., Cronquist, A. B., Silk, B. J., Jackson, K. A., O'Connor, K. A., Cosgrove, S., et al. (2013). Multistate outbreak of listeriosis associated with cantaloupe. *N. Engl. J. Med.* 369, 944–953. doi: 10.1056/NEJMoa1215837
- McDermott, P. F., Tyson, G. H., Kabera, C., Chen, Y., Li, C., Folster, J. P., et al. (2016). Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal *Salmonella*. *Antimicrob. Agents Chemother.* 60, 5515–5520. doi: 10.1128/AAC.01030-16.Address
- Mellmann, A., Andersen, P. S., Bletz, S., Friedrich, A. W., Kohl, T. A., Lilje, B., et al. (2017). High interlaboratory reproducibility and accuracy of next-generation sequencing-based bacterial genotyping in a ring trial. *J. Clin. Microbiol.* 55, 908–913. doi: 10.1128/JCM.02242-16
- Mongkolrattanothai, K., Naccache, S. N., Bender, J. M., Samayoa, E., Pham, E., Yu, G., et al. (2017). Neurobrucellosis: unexpected answer from metagenomic next-generation sequencing. *J. Pediatric Infect. Dis. Soc.* doi: 10.1093/jpids/piw066 [Epub ahead of print].
- Moore, N. E., Wang, J., Hewitt, J., Croucher, D., Williamson, D. A., Paine, S., et al. (2015). Metagenomic analysis of viruses in feces from unsolved outbreaks of gastroenteritis in humans. *J. Clin. Microbiol.* 53, 15–21. doi: 10.1128/JCM.02029-14
- Morovic, W., Hibberd, A. A., Zabel, B., Rodolphe, B., and Sthal, B. (2016). Genotyping by PCR and high-throughput sequencing of commercial probiotic products reveals composition biases. *Front. Microbiol.* 7:1747. doi: 10.3389/fmicb.2016.01747
- Naccache, S. N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., et al. (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.* 24, 1180–1192. doi: 10.1101/gr.171934.113
- Nakamura, S., Maeda, N., Miron, I. M., Yoh, M., Izutsu, K., Kataoka, C., et al. (2008). Metagenomic diagnosis of bacterial infections. *Emerg. Infect. Dis.* 14, 1784–1786. doi: 10.3201/eid1411.080589
- Nasheri, N., Petronella, N., Ronholm, J., Bidawid, S., and Corneau, N. (2017). Characterization of the genomic diversity of norovirus in linked patients using a metagenomic deep sequencing approach. *Front. Microbiol.* 8:73. doi: 10.3389/fmicb.2017.00073
- Navidad, J. F., Griswold, D. J., Gradus, M. S., and Bhattacharyya, S. (2013). Evaluation of luminex xTAG gastrointestinal pathogen analyte-specific reagents for high-throughput, simultaneous detection of bacteria, viruses, and parasites of clinical and public health importance. *J. Clin. Microbiol.* 51, 3018–3024. doi: 10.1128/JCM.00896-13
- Onori, M., Coltellà, L., Mancinelli, L., Argentieri, M., Menichella, D., Villani, A., et al. (2014). Evaluation of a multiplex PCR assay for simultaneous detection of bacterial and viral enteropathogens in stool samples of paediatric patients. *Diagn. Microbiol. Infect. Dis.* 79, 149–154. doi: 10.1016/j.diagmicrobio.2014.02.004
- Ottesen, A., Ramachandran, P., Reed, E., White, J. R., Hasan, N., Subramanian, P., et al. (2016). Enrichment dynamics of *Listeria monocytogenes* and the associated microbiome from naturally contaminated ice cream linked to a listeriosis outbreak. *BMC Microbiol.* 16:275. doi: 10.1186/s12866-016-0894-1
- Ottesen, A. R., Gonzalez, A., Bell, R., Arce, C., Rideout, S., Allard, M., et al. (2013). Co-enriching microflora associated with culture based methods to detect *Salmonella* from tomato phyllosphere. *PLoS ONE* 8:e73079. doi: 10.1371/journal.pone.0073079
- Pagotto, F., Hébert, K., and Farber, J. (2011a). “MFHPB-30. Isolation of *Listeria monocytogenes* and other *Listeria* spp. from foods and environmental samples,” in *Compendium of Analytical Methods*, Vol. 2, (Ottawa, ON: Health Products and Food Branch).
- Pagotto, F., Trottier, Y.-L., Upham, J., and Iugovaz, I. (2011b). “MFLP-74. Enumeration of *Listeria monocytogenes* in foods,” in *Compendium of Analytical Methods*, Vol. 3, (Ottawa, ON: Health Products and Food Branch).
- Park, E., Chun, J., Cha, C., Park, W., Ok, C., and Bae, J. (2012). Bacterial community analysis during fermentation of ten representative kinds of kimchi with barcoded pyrosequencing. *Food Microbiol.* 30, 197–204. doi: 10.1016/j.fm.2011.10.011
- Perez-Llarena, F. J., and Bou, G. (2016). Proteomics as a tool for studying bacterial virulence and antimicrobial resistance. *Front. Microbiol.* 7:410. doi: 10.3389/fmicb.2016.00410
- Perry, J. A., and Wright, G. D. (2014). Forces shaping the antibiotic resistome. *Bioessays* 36, 1179–1184. doi: 10.1002/bies.201400128
- Ronholm, J., Nasheri, N., Petronella, N., and Pagotto, F. (2016). Navigating microbiological food safety in the era of Q12 whole genome sequencing. *Clin. Microbiol. Rev.* 29, 837–857. doi: 10.1128/CMR.00056-16
- Ruppé, E., Baud, D., Schicklin, S., Guigon, G., and Schrenzel, J. (2016). Clinical metagenomics for the management of hospital- and healthcare-acquired pneumonia. *Fut. Microbiol.* 11, 427–439. doi: 10.2217/fmb.15.144
- Sahl, J. W., Schupp, J. M., Rasko, D. A., Colman, R. E., Foster, J. T., and Keim, P. (2015). Phylogenetically typing bacterial strains from partial SNP genotypes observed from direct sequencing of clinical specimen metagenomic data. *Genome Med.* 7, 52. doi: 10.1186/s13073-015-0176-9
- Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1186/s12915-014-0087-z
- Schellenberg, J. J., Jayaprakash, T. P., Gamage, N. W., Patterson, M. H., Vanechoutte, M., and Hill, J. E. (2016). *Gardnerella vaginalis* subgroups defined by *cpn60* sequencing and sialidase activity in isolates from Canada, Belgium and Kenya. *PLoS ONE* 11:e0146510. doi: 10.1371/journal.pone.0146510
- Schloss, P. D., and Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.* 6:229. doi: 10.1186/gb-2005-6-8-229
- Schneeberger, P. H. H., Becker, S. L., Pothier, J. F., Duffy, B., N'Goran, E. K., Beuret, C., et al. (2016). Metagenomic diagnostics for the simultaneous detection of multiple pathogens in human stool specimens from Côte d'Ivoire: a proof-of-concept study. *Infect. Genet. Evol.* 40, 389–397. doi: 10.1016/j.meegid.2015.08.044
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066
- Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* 5:209. doi: 10.3389/fpls.2014.00209
- Shea, S., Kubota, K. A., Maguire, H., Gladbach, S., Woron, A., Atkinson-Dunn, R., et al. (2017). Clinical microbiology laboratories' adoption of culture independent diagnostic tests are a threat to food-borne disease surveillance in the United States. *J. Clin. Microbiol.* 55, 10–19. doi: 10.1128/JCM.01624-16
- Smits, S. L., Schapendonk, C. M. E., Beek, J., Van Vennema, H., Schürch, A. C., Schipper, D., et al. (2014). New viruses in the Netherlands. *Emerg. Infect. Dis.* 20, 1218–1222. doi: 10.3201/eid2007.140190
- Spletstoesser, W. D., Seibold, E., Zeman, E., Trebesius, K., and Podbielski, A. (2010). Rapid differentiation of *Francisella* species and subspecies by fluorescent in situ hybridization targeting the 23S rRNA. *BMC Microbiol.* 10:72. doi: 10.1186/1471-2180-10-72
- Steyer, A., Jevšnik, M., Petrovec, M., Pokorn, M., Grosek, Š., Fratnik Steyer, A., et al. (2016). Narrowing of the diagnostic gap of acute gastroenteritis in children 0–6 years of age using a combination of classical and molecular techniques, delivers challenges in syndromic approach diagnostics. *Pediatr. Infect. Dis. J.* 35, e262–e270. doi: 10.1097/INF.0000000000001208
- Taylor, A. J., Lappi, V., Wolfgang, W. J., Lapierre, P., Palumbo, M. J., Medus, C., et al. (2015). Characterization of foodborne outbreaks of *Salmonella enterica* serovar Enteritidis with whole-genome sequencing single nucleotide polymorphism-based analysis for surveillance and outbreak detection. *J. Clin. Microbiol.* 53, 3334–3340. doi: 10.1128/JCM.01280-15
- The NIH HMP Working Group (2009). The NIH human microbiome project. *Genome Res.* 19, 2317–2323. doi: 10.1101/gr.096651.109
- Thomas, M. K., Murray, R., Flockhart, L., Pintar, K., Fazil, A., Nesbitt, A., et al. (2015). Estimates of foodborne illness-related hospitalizations and deaths in

- Canada for 30 specified pathogens and unspecified agents. *Foodborne Pathog. Dis.* 12, 820–827. doi: 10.1089/fpd.2015.1966
- Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* 4, 470–483. doi: 10.1038/nprot.2009.10
- Todd, E. C. D., Greig, J. D., Bartleson, C. A., and Michaels, B. S. (2008). Outbreaks where food workers have been implicated in the spread of foodborne disease. Part 4. Infective doses and pathogen carriage. *J. Food Prot.* 71, 2339–2373. doi: 10.4315/0362-028X-71.11.2339
- Van Belkum, A., and Dunne, W. M. (2013). Next-generation antimicrobial susceptibility testing. *J. Clin. Microbiol.* 51, 2018–2024. doi: 10.1128/JCM.00313-13
- van Duin, D., and Paterson, D. L. (2016). Multidrug-resistant bacteria in the community: trends and lessons learned. *Infect. Dis. Clin. North Am.* 30, 377–390. doi: 10.1016/j.idc.2016.02.004
- Verhoef, L., Williams, K. P., Kroneman, A., Sobral, B., van Pelt, W., and Koopmans, M. (2012). Selection of a phylogenetically informative region of the norovirus genome for outbreak linkage. *Virus Genes* 44, 8–18. doi: 10.1007/s11262-011-0673-x
- Vinjé, J. (2014). Advances in laboratory methods for detection and typing of norovirus. *J. Clin. Microbiol.* 53, 373–381. doi: 10.1128/JCM.01535-14
- Wilson, M. R., Naccache, S. N., Samayoa, E., Biagtan, M., Bashir, H., Yu, G., et al. (2014). Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N. Engl. J. Med.* 370, 2408–2417. doi: 10.1056/NEJMoa1401268
- Won, Y. J., Park, J. W., Han, S. H., Cho, H. G., Kang, L. H., Lee, S. G., et al. (2013). Full-genomic analysis of a human norovirus recombinant GII.12/13 novel strain isolated from South Korea. *PLoS ONE* 8:e85063. doi: 10.1371/journal.pone.0085063
- Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46
- Zagordi, O., Bhattacharya, A., Eriksson, N., and Beerenwinkel, N. (2011). ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 12:119. doi: 10.1186/1471-2105-12-119
- Zhao, X., Lin, C.-W., Wang, J., and Oh, D. H. (2014). Advances in rapid detection methods for foodborne pathogens. *J. Microbiol. Biotechnol.* 24, 297–312. doi: 10.4014/jmb.1310.10013
- Zhou, Y., Wylie, K., Ei Feghaly, R. E., Mihindukulasuriya, K., Elward, A., Haslam, D. B., et al. (2016). Metagenomic approach for identification of the pathogens associated with diarrhea in stool specimens. *J. Clin. Microbiol.* 54, 368–375. doi: 10.1128/JCM.01965-15
- Zilelidou, E., Karmiri, C., Zoumpopoulou, G., Mavrogonatou, E., Kletsas, D., and Tsakalidou, E. (2016). *Listeria monocytogenes* strains underrepresented during selective enrichment with an iso method might dominate during passage through simulated gastric fluid and *in vitro* infection of caco-2 cells. *Appl. Environ. Microbiol.* 82, 6846–6858. doi: 10.1128/AEM.02120-16

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Forbes, Knox, Ronholm, Pagotto and Reimer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Metagenomic Sequencing for Surveillance of Food- and Waterborne Viral Diseases

David F. Nieuwenhuijse and Marion P. G. Koopmans*

Department of Viroscience, Erasmus Medical Center, Rotterdam, Netherlands

A plethora of viruses can be transmitted by the food- and waterborne route. However, their recognition is challenging because of the variety of viruses, heterogeneity of symptoms, the lack of awareness of clinicians, and limited surveillance efforts. Classical food- and waterborne viral disease outbreaks are mainly caused by caliciviruses, but the source of the virus is often not known and the foodborne mode of transmission is difficult to discriminate from human-to-human transmission. Atypical food- and waterborne viral disease can be caused by viruses such as hepatitis A and hepatitis E. In addition, a source of novel emerging viruses with a potential to spread via the food- and waterborne route is the repeated interaction of humans with wildlife. Wildlife-to-human adaptation may give rise to self-limiting outbreaks in some cases, but when fully adjusted to the human host can be devastating. Metagenomic sequencing has been investigated as a promising solution for surveillance purposes as it detects all viruses in a single protocol, delivers additional genomic information for outbreak tracing, and detects novel unknown viruses. Nevertheless, several issues must be addressed to apply metagenomic sequencing in surveillance. First, sample preparation is difficult since the genomic material of viruses is generally overshadowed by host- and bacterial genomes. Second, several data analysis issues hamper the efficient, robust, and automated processing of metagenomic data. Third, interpretation of metagenomic data is hard, because of the lack of general knowledge of the virome in the food chain and the environment. Further developments in virus-specific nucleic acid extraction methods, bioinformatic data processing applications, and unifying data visualization tools are needed to gain insightful surveillance knowledge from suspect food samples.

OPEN ACCESS

Edited by:

Jennifer Ronholm,
McGill University, Canada

Reviewed by:

Learn-Han Lee,
Monash University Malaysia Campus,
Malaysia
Tineke H. Jones,
Agriculture and Agriculture-Food
Canada, Canada

*Correspondence:

Marion P. G. Koopmans
m.koopmans@erasmusmc.nl

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 16 November 2016

Accepted: 01 February 2017

Published: 15 February 2017

Citation:

Nieuwenhuijse DF and
Koopmans MPG (2017)
Metagenomic Sequencing
for Surveillance of Food-
and Waterborne Viral Diseases.
Front. Microbiol. 8:230.
doi: 10.3389/fmicb.2017.00230

INTRODUCTION

Transmission via the food- and waterborne route is a common mode of spread of a wide range of viruses. Many commonly recognized food- and waterborne infections are caused by viruses that are transmitted by the fecal-oral route. Particularly caliciviruses (norovirus, sapovirus) can cause diarrhea and vomiting and less commonly astroviruses, rotaviruses, and adenoviruses (Newell et al., 2010). Other viruses cause symptoms resulting from extra-intestinal spread, like hepatitis A (HAV), and hepatitis E (HEV). High levels of viral shedding through stool and vomit lead to dispersal in the environment. Moreover, the stability of many food- and waterborne viruses allows

for prolonged persistence in the environment. Food- and water associated transmission is also suspected to enhance the spread and emergence of zoonotic viruses (e.g., Middle East Respiratory Syndrome-coronavirus and Nipah virus) and facilitates the occurrence of zoonotic events through the handling of bushmeat (Ebola virus) (Wolfe et al., 2005; European Food Safety Authority, 2014; Mann et al., 2015).

Challenges of detecting viruses transmitted by the food- and waterborne route are their diversity and the frequent secondary person-to-person transmissions, which may mask an initial food- or waterborne introduction. In addition, there is a lack of awareness among clinicians (Beersma et al., 2012), as the symptoms caused by foodborne viruses are not specific to the viruses causing the illness. Furthermore, there is limited coverage in surveillance of food- and waterborne viral disease, hampering detecting and tracing (Ahmed et al., 2014; Verhoef et al., 2015).

In the past years, high-throughput sequencing technologies have increased the ability to measure genomic material from diverse samples tremendously. These methods will most likely continue to improve in the future (Aarestrup et al., 2012). Specifically, metagenomic analysis using untargeted sequencing has received a lot of attention, because the high throughput of current sequencing technologies has made it possible to obtain multiple high coverage genomes from highly complex samples (Cotten et al., 2014; Smits et al., 2015). Even though it is still a developing field, metagenomics is starting to become mature enough for applications outside of the research environment.

With the development of multiplex real-time polymerase chain reaction (RT-PCR) protocols came the realization that unraveling etiologies of main disease syndromes is more complex than previously recognized. This led to questions about the detection of viruses for which the role as causes of illness remains to be evaluated, the importance of co-infections and recognition of less common disease etiologies (Binnicker, 2015). Similarly, high throughput metagenomic sequencing broadens the scope of detectable viruses, which, apart from making it more complex, make us further understand the role of viruses in health and disease. The biggest promise, however, is that of routine application of metagenomic sequencing in diagnostic context, facilitating viral detection and offering huge potential for tracing of viruses in (foodborne) outbreaks.

RECOGNIZING FOOD- AND WATERBORNE VIRAL DISEASE

Given the number of different viral pathogens potentially associated with food- and waterborne transmission their detection has not been straightforward. Partly because many of these pathogens lack cell culture systems that are sensitive and robust enough for application in routine settings (Amar et al., 2007). The entry point for disease-based surveillance of viruses spreading by food and water is the reporting of patients presenting to a clinician. However, patients only present themselves in case of a severe symptomatic infection, or in case self-help is not sufficient. Mild symptoms are therefore generally not registered creating a bias in surveillance. This phenomenon

is captured in the surveillance pyramid (**Figure 1**), and the full extent of disease can only be captured through epidemiological studies addressing incidence and etiology at community level coupled with severity of a range of enteric pathogens (Sethi et al., 1999; de Wit et al., 2001; Tam et al., 2012). Additionally, it is challenging to distinguish between foodborne outbreaks and outbreaks caused by direct contact between humans. Classic clinical symptoms of foodborne disease vary, ranging from diarrhea and vomiting to abdominal cramps and general malaise, which makes it hard for clinicians to pinpoint the exact causative agent. This leads to misdiagnosis if the diagnostic workup is selective, and if there are no obvious signs of food-related exposure (Beersma et al., 2012). Moreover, heterogeneity in clinical interpretation can be caused by host factors, such as differences in the expression of histo-blood-group antigens that are receptors for rota- and noroviruses (Payne et al., 2015; de Graaf et al., 2016). Susceptibility to fecal-orally transmitted viruses may also be influenced by the established microbiome and virome in the host population, of which the prior is shown to differ between different locations and age groups (Yatsunenko et al., 2012). It is reasonable to think that the differences in the gut environment are more pronounced between countries with larger social and economic differences such as first and third world countries, which often differ in their resident pathogens (Ott et al., 2012; Yatsunenko et al., 2012; Hay et al., 2013). The role of the gut virome, in addition to the gut microbiome, is a relatively new concept and has been described as potentially

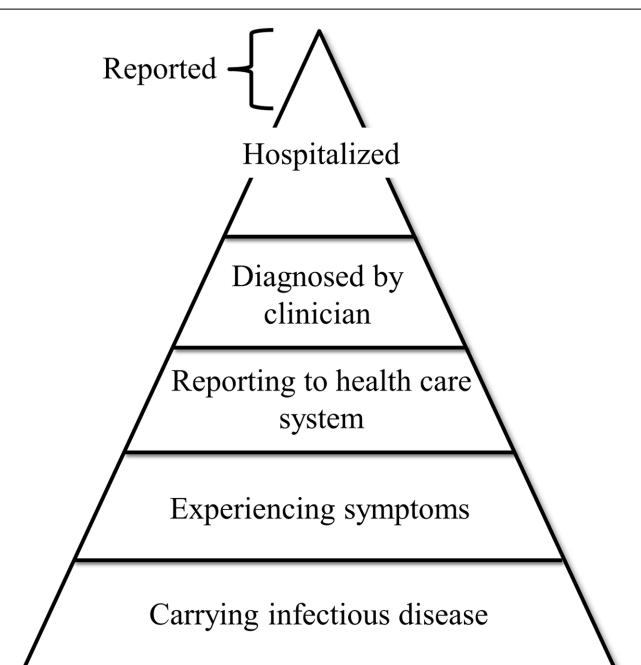


FIGURE 1 | Schematic representation of the phenomenon known as the “surveillance pyramid”. Layers represent different categories of infected individuals. Width of the layers represents the estimated number of individuals in that category. As indicated, individuals reported by surveillance programs generally originate from the hospitalized category.

having influence on gut health and therefore expression of disease (Cadwell, 2015). Because of under and miss-diagnoses, clinical surveillance likely only captures the tip of the iceberg of food- and waterborne viral disease cases.

DETECTION OF FOOD- AND WATERBORNE VIRAL DISEASE OUTBREAKS

In cases where a cluster of patients with similar symptoms presents itself, there can be an investigation to look for epidemiological clues of the link between the cases. Additional information is garnered from the use of viral genome sequencing, making it possible to track origins of outbreaks, and to estimate how much of the observed human disease is attributable to foodborne infection by computerized linking of epidemiologic data to aligned viral genomic sequences (Verhoef et al., 2011). However, often the original source or evidence of it being food- or waterborne cannot be found, which means that outbreaks often are merely registered. Of the 941 viral disease outbreaks reported as foodborne in the joint ECDC-EFSA surveillance report of 2015, only 9.1% had robust evidence of food- or waterborne transmission (Eurosurveillance editorial team, 2015). Routine application of genotyping of HAV in newly diagnosed cases quadrupled the number of cases in which food was the most likely source of infection a 3 year enhanced surveillance study in The Netherlands, but this is not commonly done (Petrignani et al., 2014). In an investigation of 1794 food- and waterborne outbreaks in Korea, roughly 75% of the outbreaks reported in schools and public restaurants were attributed to an unknown origin (Moon et al., 2014). Availability and costs of molecular testing combined with sequencing, additional to the limited success of virus detection in food products, are likely further limiting their use in food and water surveillance. This is demonstrated by the fact that formal confirmation of a viral outbreak associated with food- and waterborne transmission still requires extensive epidemiological analysis or confirmation of a virus in the infected individual, or both (EFSA, 2016). However, due to the increase of genomic information of viruses, sequence data is increasingly used to support and strengthen outbreak investigations. Nevertheless, the surveillance programs for these viruses in the human food chain is limited, in contrast with the American CDC¹ and the European ECDC² surveillance programs for bacteria and parasitic pathogens causing food- and waterborne diseases (Deng et al., 2016) and does not have widespread coverage. As an example, to comply with European food safety regulations, shellfish, a well-known source of foodborne pathogens, need to be tested for enteric bacteria. However, it has been well documented that shellfish that pass quality control based on bacterial counts may still contain human pathogenic viruses (Rodriguez-Manzano et al., 2014). To be able to recognize food- waterborne viral disease outbreaks and

stop underestimation of its disease burden there should be innovations in the current foodborne surveillance system.

CLASSICAL VIRUSES ASSOCIATED WITH FOOD- AND WATERBORNE DISEASES

Although the list of viruses causing acute gastroenteritis is long, norovirus ranks among the top causes of diarrheal disease (Ahmed et al., 2014). Reporting of outbreaks suggests that the food- and waterborne disease transmission route is relatively rare, but provides an underestimate, bearing in mind that it may be hard to recognize a food- and waterborne transmission route in community-acquired diarrheal disease. To quantify the burden of all diarrheal disease attributable to foodborne transmission, the World Health Organization commissioned a study that combined data from surveillance and exhaustive literature reviews with a systematic approach to calculation of the fraction of disease attributable to food contamination (Havelaar et al., 2015). This ranked the burden of norovirus illness among the top causes of foodborne disease, along with *Campylobacter*, and listed HAV associated disease among other significant causes of foodborne disease, along with *Salmonella* and *Taenia solium*.

For bacterial foodborne pathogens, the analysis of systematically collected surveillance data has been used as the basis of attribution analysis (Pires et al., 2009). A popular approach has been to quantify the proportion of foodborne disease of humans to their likely origin, by comparing diversity of strains found in human disease outbreaks with that found in animal and environmental reservoirs (Hald et al., 2007). While this model does not allow estimating the foodborne disease where food is a vehicle for person-to-person transmission, which is common for noroviruses, it has been used with some success to quantify the contribution of foodborne viral disease stemming from environmentally contaminated food (e.g., associated with shellfish; (Verhoef et al., 2015)). This builds from the observation that there is a large discrepancy between the norovirus variants in clinical settings and environmental samples (Tao et al., 2015; Kazama et al., 2016). Norovirus GII.4, found in clinical setting, is generally related to person-to-person transmission, however, several other norovirus genotypes and genogroups were found in environmental samples in the same area. However, food associated acute gastroenteritis is not limited to norovirus infections. In a large retrospective study of oyster-related acute gastroenteritis outbreaks in Osaka City in Japan 30.7% of the cases were attributed to other pathogens such aichivirus, astrovirus, sapovirus rotavirus A, and enteroviruses (Iritani et al., 2014). Furthermore, outbreaks can be caused by a mixture of these viruses and viral variants (Wang et al., 2015).

OTHER VIRUSES TRANSMITTED VIA THE FOOD- AND WATERBORNE ROUTE

Apart from viruses causing gastroenteritis, there are viruses causing food- and waterborne diseases that are associated with

¹http://ecdc.europa.eu/en/healthtopics/food_and_waterborne_disease/surveillance/Pages/index.aspx

²<http://www.cdc.gov/ncezid/dfwed/keyprograms/surveillance.html>

a variety of other syndromes. The second most common disease syndrome is hepatitis, caused by HAV, a fecal-orally transmitted virus (Havelaar et al., 2015). By decreasing natural exposure in regions with low endemicity, the susceptibility of the population for outbreaks of HAV disease in these regions is increasing (Newell et al., 2010). Because of increased globalization, contamination of food products by viruses prevalent in food producing regions can increase the risk of outbreaks in these regions. Several outbreaks of HAV infection have been reported in recent years both in the USA and Europe (Gossner et al., 2014). Most of these outbreaks could be identified as foodborne infections after intense investigations (Bruni et al., 2016). Especially fresh (imported) food products (e.g., fresh frozen berries, pomegranate seeds, and sun-dried tomatoes) have been identified as sources of the virus (Gossner et al., 2014; Tavoschi et al., 2015). Tracking the foodborne source of infection is challenging for HAV, because of an underestimation of the contribution of food as a source of infection due to the long incubation period in infected individuals (Petrignani et al., 2014).

Another foodborne virus gaining increased attention is zoonotic HEV, associated with genotype 3 and 4 HEV. HEV is widespread in commercially held pigs, as well as in wild pigs, and deer (Guillois et al., 2016). Human disease with genotype 3 HEV is increasingly recognized, but in the large majority of the cases the source of the virus is unknown (Lewis et al., 2010). There is clear evidence that food can be a source of zoonotic HEV infections. Outbreaks that have been confirmed to be caused by foodborne transmission of the virus by consumption of wild meat from boar, deer, and rabbit (Tei et al., 2003; Izopet et al., 2012; Guillois et al., 2016). Several studies have shown the zoonotic potential of HEV from pigs (Teixeira et al., 2016), HEV can also be readily detected in pork products such as dried meats and liver sausages (Di Bartolo et al., 2015). A large proportion of food-related HEV infections, however, does not lead to hospitalization of the patient, leading to under-reporting and unrecognized risk and burden of the disease (Guillois et al., 2016).

Beside viruses circulating in livestock, wildlife has the potential to be a large reservoir of unknown zoonotic viruses. Hunting, trading, preparing, and consuming so-called “bushmeat” is one of the routes by which novel viruses can be introduced into the human population (Karesh and Noble, 2009). It may be difficult to disentangle foodborne infection from direct zoonotic exposure, but it is important to consider local practices before ruling out food as a source of human infection. A special example are the occasional introductions of Nipah viruses from bats into humans through contamination of date palm sap which is collected in open containers to which bats that harbor these viruses have access (Rahman et al., 2012). Not proven but certainly interesting is the practice of drinking unprocessed camel urine which may contain MERS coronavirus, a practice that came to light during the investigations into sources of MERS coronavirus infection in humans (Funk et al., 2016). Even if limited in scale, small foodborne infections, originating from human-wildlife interaction, constitute as many incidents potentially pushing wildlife viruses to become human-to-human transmissible (Wolfe et al., 2005; Islam et al., 2016). In the cases of Monkeypox and Nipah this only led to small epidemics,

but when the virus is well adapted to spread from human to human this can lead to larger outbreaks, as seen during the Ebola crisis in 2015 (Wolfe et al., 2005; Mann et al., 2015). Continuing deforestation, increasing population and continued trade of bushmeat brings more humans in contact with wildlife and increases the risk of zoonosis (Karesh and Noble, 2009). Urbanization and globalization of travel and trade provides ample and increasing opportunity for further spread. Therefore, even anecdotal zoonotic introductions may constitute a public health risk, and ideally should be investigated in conjunction with the animals these humans were exposed to. As the ability to spread between humans is a key property for successful further spread, enhancing the capacity to investigate clusters of disease (in humans and animals) is important (McCloskey et al., 2014).

UNKNOWN FECAL-ORAL PASSENGERS

Bacteriophages, although not directly pathogenic to humans, could play a role in human health and disease by influencing the gut microbiome. Sequencing data from human gut samples presents a large diversity of bacteriophages in the human gut (Reyes et al., 2012). In addition to bacteriophages, untargeted sequencing of sewage samples has shown the presence of large quantities of different plant viruses (Zhang et al., 2006). Because of the presence of numerous infectious plant virus particles in human fecal waste there is ongoing research on the effect of these viruses in human health and disease (Colson et al., 2010). Similarly, there is ongoing research into the impact of bacteriophages on human health through their modulating effect on the gut microbiome (Reyes et al., 2012), and thereby, gut immunity (Honda and Littman, 2016). In what way bacteriophages protect or expose the human gut to bacterial or viral pathogens has yet to be further investigated. However, using metagenomic sequencing it will at least be possible to recognize the presence of unknown fecal-oral passengers.

METAGENOMICS FOR FOOD- AND WATERBORNE VIRAL DISEASE SURVEILLANCE

Metagenomics is a term used for experiments in which all nucleic acids in a certain sample are sequenced. For bacteria, historically, the diversity of a sample used to be expressed by performing phylogenetic analyses based on 16S ribosomal RNA (Handelsman, 2004). However, since viruses lack such a universally conserved motif, viral metagenomics refers to the attempt to recover full and partial genomes of all viruses present in the sample. Viral metagenomic analysis protocols generally start with procedures to remove host and bacterial cells followed by nuclease treatment to remove free nucleic acids. Often, the remaining nucleic acids are amplified using randomly primed (RT-) PCR and finally sequenced using high-throughput sequencing technology. Viral metagenomics has great potential in surveillance of viruses in the global food chain because of its

sensitivity, broad detection range, and detailed information of the detected virus (Aarestrup et al., 2012).

Environmental Surveillance

Metagenomic sequencing has already been used in the sampling of the world's oceans to estimate the global viral diversity (Hingamp et al., 2013). Similarly, metagenomics can be used in environments associated with viruses spread via the food- and waterborne route (**Figure 2A**), which gives an overview of all these viruses and circumvents the mentioned sampling biases. The potential of such an approach for food-related purposes was exemplified by Hellmér et al. (2014) who conducted a multi-species viral surveillance study and, albeit not metagenomic sequencing based, were able to detect several food- and waterborne viruses in sewage. Interestingly, norovirus and HAV, detected in sewage, could be related to hospitalized patients diagnosed with the viral infection in the catchment area of the sewage system. Moreover, they detected a peak in the level of norovirus several weeks before the outbreak was reported in the hospital in that area (Hellmér et al., 2014). This demonstrates the potential power of shifting the scope of surveillance of food- and waterborne viruses from the hospital to the environment. Untargeted metagenomic sequencing has been shown to be able to capture a multitude of viruses in sewage samples in several studies. Moreover, comparison between sewage viromes from Nigeria, Nepal, Bangkok, and California, four geographically distant locations, showed distinct differences in the subsets of detected human viruses (Ng et al., 2012). Interestingly, the average sequence similarity between

the reference sequences stored on the NCBI GenBank and the human viruses detected in the samples from California was higher than those from the other locations. This may indicate a bias towards American viruses in view of human virus diversity in this database (Ng et al., 2012). A study that looked at viruses from sewage capable to infect human epithelial cells was able to detect a large number of bacteriophages and several different species of the *Polyomaviridae*, *Picornaviridae*, and *Papillomaviridae* viral families.(Aw et al., 2014). Another more recent evaluation of untargeted metagenomic sequencing for surveillance purposes retrieved full genomes of Adeno-associated virus-2 as the most prominent mammalian virus in the sample. This virus is generally not associated with any pathology and cannot be grown in cell cultures, possibly underestimating its role in diarrheal disease (Furtak et al., 2016). A striking fact of these studies is the number of sequencing reads that are found that share no sequence similarity with current reference databases. Percentages of unmapped sequences range from 37 to 66% (Cantalupo et al., 2011; Ng et al., 2012). Whether these sequences represent novel viruses that can be transmitted via the food- and waterborne route remains to be determined. Nevertheless, these preliminary studies show the potential of untargeted metagenomic sequencing to detect novel and known human pathogens. Sampling a larger variety of locations, performing longitudinal studies of the same environment and deeper sequencing will provide more information on what environmental metagenomic sequencing can contribute to the monitoring of viral trends and viral diversity.

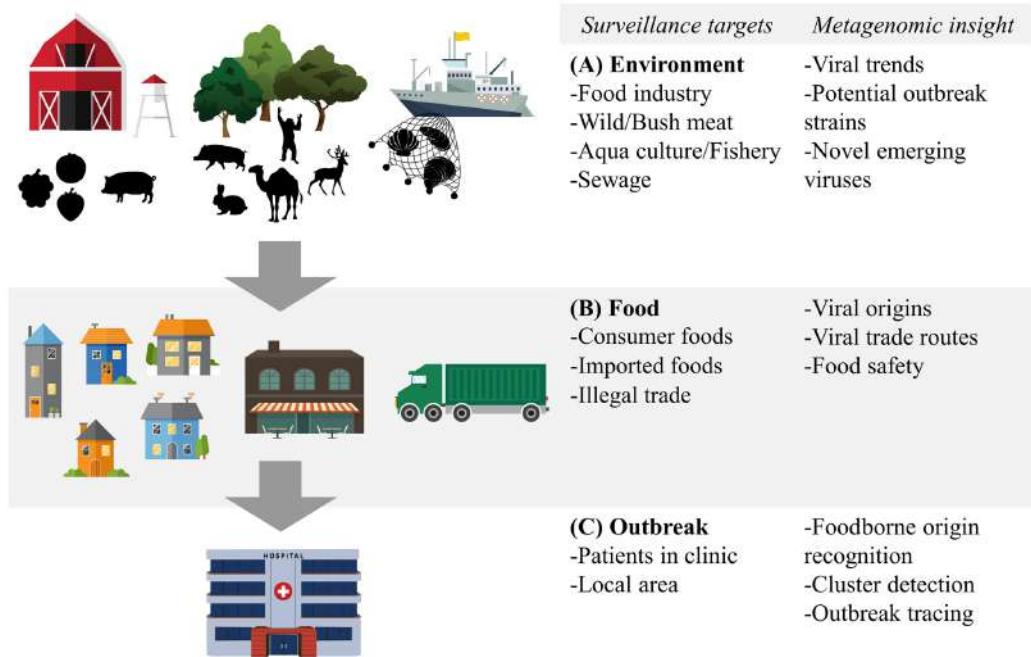


FIGURE 2 | Food- and waterborne viral surveillance targets for metagenomic sequencing approaches. **(A)** Environmental surveillance of food industry, wild meat and bushmeat habitat, and aquaculture and fishery environment. **(B)** Food surveillance of consumer and imported foods, including illegally imported foods. **(C)** Surveillance of food- and waterborne outbreaks, in clinic and locally. Potential of metagenomic sequencing based surveillance is listed next to each category.

Food Surveillance

Analogous to the environment in which it has been produced, food itself can benefit from metagenomic surveillance. Food contamination in combination with international trade, changing eating habits and food processing practices all contribute to the spread of food- and waterborne viruses and making food itself a valuable target of metagenomic surveillance (**Figure 2B**). Sentinel screening of imported foods, especially risk foods such as fresh fruits and vegetables, dried meats and seafood, could prevent foodborne viral outbreaks such as the international HAV outbreak in Europe from 2012 to 2013 (Severi et al., 2015). Successful application of metagenomic sequencing of viruses has been shown in a study isolating viruses in the family of *Reoviridae* and *Picobirnaviridae* from field-grown lettuce (Aw et al., 2014).

Apart from legal trade, illegal import of food products, such as bushmeat, could also be screened. Untargeted metagenomic sequencing is especially suited for these types of screenings, as the origin and the potential viral content of these samples are often completely unknown. In one example, metagenomic sequencing was performed on bushmeat seized by the customs officers of a French airport. Although no viruses with a potential threat to human health could be detected (Temmam et al., 2016), these initial attempts should be looked at as potentially interesting surveillance approaches, given that relatively large quantities of raw bushmeat are estimated to enter Europe and the Americas annually (Mann et al., 2015).

Another source of known and potentially unknown foodborne disease-causing viruses are shellfish. Mainly the consumption of oysters is associated with foodborne outbreaks (Bellou et al., 2013). However, oysters, cockles, and clams have been shown to accumulate norovirus, sapovirus, and HAV (Benabbas et al., 2013). To our knowledge, there are no published studies performing untargeted virome sequencing of these shellfish. Surveillance by metagenomic sequencing can be beneficial for aquaculture, also for monitoring seafood health, as in aquaculture, large numbers of animals are kept in a confined environment for an extended period, increasing opportunities for the spread of infections. Cultivated fish and other sources of seafood can be infected with a wide variety of viruses (Alavandi and Poornima, 2012).

Outbreak Surveillance

One of the main promises of surveillance using metagenomic sequencing is that of concomitant clinical application (diagnosis of patients) and public health application (typing and cluster analysis to trace of food- and waterborne outbreaks) (**Figure 2C**). Using metagenomic sequencing, the effort of detecting and genotyping of a virus can be combined to trace an outbreak, regardless of prior knowledge of the virus, provided the data is analyzed in combination with relevant metadata. The use of this integrative approach has been demonstrated in an investigation of a hospital outbreak of human parainfluenza virus, which was investigated using high-throughput metagenomic sequencing (Greninger et al., 2016). Both the detection of the virus, the diagnosis of the disease and the establishment of viral clusters and transmission routes could be derived from the

metagenomic sequencing data. A similar approach should enable investigation of viruses related to food- or waterborne diseases and distinguishing between a food- and waterborne and a person-to-person transmission route. In such investigations, speed is of the utmost importance, therefore on-site sequencing strategies, enabled by novel portable sequencing platforms such as the Oxford Nanopore MinION (Hoelen et al., 2016), have potential in fast local outbreak detection and disease monitoring (Arias et al., 2016). Recent reports have shown potential in metagenomic detection of hepatitis C, chikungunya, Ebola and Zika virus in hospital settings (Greninger et al., 2015; Sardi et al., 2016). The development of on-site sequencing technology is still in its infancy, however, and it remains to be investigated if food-related viral outbreaks will be traceable and can deliver whole-genome based viral dynamics analysis analogous to the investigation of the Ebola outbreak of 2014 (Gire et al., 2014; Quick et al., 2016). However, the same on-site technology has been shown to be beneficial in tracing foodborne salmonella (Quick et al., 2015). Aspects of current on-site sequencing technologies that need to be improved for viral metagenomic sequencing are the limited throughput and sequence quality, which limit the detection of low-level viral genomes and minor variants. Nevertheless, the use of near-real-time sequencing of Ebola and Zika during the recent outbreaks has received a lot of attention and has shown that the technology works.

CHALLENGES IN SAMPLE PREPARATION AND SEQUENCING

The routine application of metagenomic sequencing for clinical diagnosis and surveillance is dependent costs versus performance criteria such as speed, reliability, and comparability of results with those of reference methods. Improvements are necessary in the standardization and speedup of sample preparation, sequencing and data analysis for clinical and public health application. A recent study has shown the potential of fast whole-genome sequence based epidemiological tracing in the recent Ebola outbreak (Arias et al., 2016). However, specific primers were used to target the Ebola genome, which is different from a metagenomic sequencing approach.

The developments and different choices of sequencing technology make it difficult to decide how to standardize routine diagnostics and surveillance protocols. Studies that directly compare platforms help in this decision making process. Two studies compared the Illumina MiSeq, Roche-454 titanium, Ion Torrent PGM, and PacBio RS platforms (Quail et al., 2012; Frey et al., 2014). For viral metagenomics application, the Illumina and the Ion Torrent platform seem to outperform the other two platforms based on their relatively low cost per giga base output. Between these two systems, the main tradeoff is the sequencing time versus the sequencing read output. The high volume of sequencing reads produced by the Illumina platform, in a longer timeframe, increases the chance that a lowly abundant viral genome is sufficiently covered, which makes it more suitable for metagenomics of complex samples. The Ion Torrent platform delivers a smaller number of reads in a smaller timeframe, which

is beneficial when a timely result is necessary, and low level presence of viruses is disregarded, for instance in diagnostic settings.

Novel approaches, such as the MinION nanopore sequencer, increase speed and depth of coverage at the cost of sequence error rate. A comparison of a metagenomics approach using the MinION nanopore or the Illumina MiSeq sequencer reports a sample-to-result time of 6 h for a MinION nanopore setup compared to 20 h using an Illumina MiSeq setup (Greninger et al., 2015). Despite their reported successes in identification of viruses, the reported error rate of 10 to 60 percent impedes high resolution sequence classification at low genome coverage, or the use of sequence data for reliable source-tracking. It does allow very rapid virus classification in cases where low coverage suffices, or at high viral titers (Hoenen et al., 2016; Quick et al., 2016). However, performance of the MinION sequencing platform remains to be tested at lower virus titers and with more complex samples which are generally encountered in surveillance and clinical settings (food, feces, sewage).

To increase the viral specific output of metagenomics sequencing approaches sample preparation methods can be used to reduce non-viral genomic material or specifically select for viruses. Approaches that are being investigated, range from different extraction protocols (Cotten et al., 2014; Conceição-Neto et al., 2015) to using a virome specific capturing chip (Briese et al., 2015) or blood-derived antibodies to capture viral particles (Oude Munnink et al., 2013). Paradoxically, however, the sequencing capacity of high throughput metagenomic sequencing is sensitive enough to pick up contaminants from the lab reagents, or from previous experiments (Gruber, 2015). These pose a challenge to the interpretation of metagenomic data. To limit contamination, laboratories in which samples are processed are often separated from those in which nucleic acids are amplified and equipment is UV treated and cleaned with bleach. Additionally, alternating the sample-specific DNA barcodes in multiplex sequencing experiments reduces contamination from previous runs. Nevertheless, it is recommended to include both negative control samples, which have been processed similarly, but are believed to contain no viruses, and positive control samples, that contain known quantities of a variety of viruses (Lusk, 2014). Alternatively, bioinformatics tools such as DeconSeq (Schmieder and Edwards, 2011), have been developed to check for signals of regularly found lab contamination in the sequencing data. In conclusion, as contamination of samples and equipment may not be avoidable, its likelihood should be taken in consideration when using metagenomic sequencing technology for food-related surveillance applications.

DIFFICULTIES OF METAGENOMIC DATA ANALYSIS

Aside from lab-based technical difficulties, there are several challenges concerning data analysis of metagenomic sequencing experiments. First, due to the high and increasing read output of sequencing machines, data analysis of high throughput sequencing projects generally requires strong computational

infrastructure, which, can require large investments and technological expertise (Spjuth et al., 2016). However, metagenomic data analysis tools have been improving, optimizing the ratio between computing resources needed and their speed and accuracy. Sequence annotation tools based on k-mer lookup tables such as UBLAST (Edgar, 2010), Kraken (Wood et al., 2014), Kaiju (Menzel et al., 2016), and Diamond (Buchfink et al., 2014) have increased the speed of sequence assignment to reference database with several orders of magnitude, while requiring relatively modest processing power.

Second, the assembly of millions of genomic fragments into 1000s of different individual genomes is a daunting task. Historically, short-read assemblers were developed and optimized to assemble a single genome out of a set of sequencing reads. These assemblers are therefore not suited for the reconstruction of metagenomes and are prone to creating synthetic chimeric genomes (Vázquez-Castellanos et al., 2014). Various assemblers have since been developed specifically aimed at metagenome assembly, like MetaSPAdes (Nurk et al., 2016), Ray-Meta (Boisvert et al., 2012), MetAMOS (Treangen et al., 2013), MetaVelvet (Afiahayati et al., 2015), and IDBA-UD (Peng et al., 2012). Nevertheless, metagenome assembly is still a challenging task, often requiring manual editing to resolve miss-assemblies.

Third, assigning all assembled genomes to a reference genome is hampered by miss-annotations and incomplete reference databases. One example is “non-A, non-B hepatitis virus”, a sequence present in the NCBI GenBank, which was miss-annotated and the sequence was shown to belong to a bacteriophage (Cantalupo et al., 2011). The volume of sequencing databases is increasing rapidly, however, sequence annotations and metadata are of varying levels of quality and the speed of analysis decreases with increasing reference datasets. Therefore, there is a tradeoff between the rate of success of annotation of a sequence against a smaller curated reference dataset, and reliability of annotation using a large reference database with less-well curated annotation data.

Sequence homology of multiple reference genomes can lead to the spurious assignment of sequencing reads to one of these genomes. An example of the impact of spurious read annotation was the alleged detection of genomic material of *Yersinia pestis* in the New York subway system. Further inspection showed that the reads mapping to *Yersinia pestis* could have mapped with similar likelihood to other bacterial species (Afshinnekoo et al., 2015). Such miss-annotations of metagenomic sequences need to be anticipated and carefully addressed before using metagenomics in surveillance and diagnostic applications.

METAGENOMIC DATA INTERPRETATION

The final challenge of metagenomic sequencing based surveillance is the interpretation of the annotated sequences. There is still little knowledge of the presence and dynamics of viruses in the environment and the food chain, which is of influence on the interpretation of food- and waterborne viral surveillance samples. Various factors are expected to influence

the virome, and without knowledge of the typical viral content of a sample, the relevance of the detection of a virus is hard to determine. An example of this is a study showing a large discrepancy between the levels of HAV genotypes detected in sewage samples compared to the genotype infecting patients in the clinic in the same time (La Rosa et al., 2014). A potential sampling bias and asymptomatic shedding of one of the variants was proposed as an explanation of the discrepancy. However, this shows that a lack of knowledge of viral diversity in a population under surveillance could potentially lead to wrong conclusions in environmental surveillance studies.

Detection of a virus by molecular methods relies on intact genomic material of a virus being present in the sample. However, the relationship between the infectivity of a detected virus and the detection of a fragment of its genome is not unambiguous. Apart from intrinsic virus characteristics, infectivity and detection of a virus depends on its stability in the sample matrix (Cook and Rze, 2004) and during sample preprocessing steps (Conceição-Neto et al., 2015). Similarly, the detection of a virus using untargeted metagenomic sequencing does not confirm its infectivity. Cell culture based infectivity assays are the golden standard to determine virus infectivity, these methods are, however, not scalable and many viruses cannot be cultured *in vitro* (Hamza et al., 2011). High genome coverage combined with close sequence identity to a viral reference genome with a known pathogenic phenotype are currently the strongest links between metagenomic sequencing data and disease etiology. Nevertheless, currently employed PCR based methods, which are based on genome fragment detection, suffer from the same limitations (D'Agostino et al., 2011).

It is becoming increasingly clear that integration of different data sources and experimental results is crucial for the interpretation of metagenomic sequencing experiments. Therefore, browsing of these data and visualization of relationships between genome datasets and metadata should be facilitated. In the recent years, interactive web-based data browsing and visualization tools have increased in popularity to facilitate the interaction with and the browsing through highly complicated data in a user-friendly manner. Further development of tools that facilitate interaction with and visualization of metagenomic sequencing results, such as Kronatools (Ondov et al., 2011) and Taxonomer (Flygare et al., 2016), and frameworks for so-called data analysis “dashboards”^{3,4,5}, should make the interpretation of metagenomics experiments easier in the future.

³ <http://shiny.rstudio.com/>

⁴ <https://plot.ly/>

⁵ <http://jupyter.org/>

REFERENCES

- Aarestrup, F. M., Brown, E. W., Detter, C., Gerner-Smidt, P., Gilmour, M. W., Harmsen, D., et al. (2012). Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. *Emerg. Infect. Dis.* 18:e1. doi: 10.3201/eid/1811.120453
- Afiahayati, R., Sato, K., Sakakibara, Y., Robertson, D. L., Prosperi, M., Afiahayati, K., et al. (2015). MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing

CONCLUSION

In our current society, there is much attention for the diseases that are causing occasional outbreaks. However, there are multiple strong signs that there are viruses hiding below the radar, due to a focus on viruses with direct clinical impact. As such, the disease burden of food- and waterborne viral infections is mainly recorded in outbreaks, signified by severe symptoms and hospitalization. However, it is estimated that the large abundance of viral infections causing mild symptoms, and thus not being recorded, carry a large portion of the global food- and waterborne disease burden. Moreover, this disease burden is expanded by the consequential infections and outbreaks of these viruses in susceptible populations. Global food trading, diversification of food sources and interactions with animals and other reservoirs of food- and waterborne disease related viruses complicate the capability of investigators to detect the original source and to determine the transmission pattern of viruses causing foodborne outbreaks. Therefore, surveillance efforts should look to metagenomic sequencing technologies, bioinformatics analysis tools and data sharing initiatives to get a more realistic insight in the global burden of food- and waterborne viral disease, and to make informed decisions on how to reduce this burden.

AUTHOR CONTRIBUTIONS

DN and MK designed the focus of the review, DN did the literature search and drafted the manuscript, MK reviewed and revised.

FUNDING

European Union's Horizon 2020 research and innovation program under grant agreement No. 643476 (COMPARE). ZonMW TOP project 91213058.

ACKNOWLEDGMENTS

We thank Matt Cotten, Miranda de Graaf, Bas Oude Munnink and My Phan for their critical review of the manuscript and valuable discussions. We would like to give credit to www.vecteezy.com, who provided the vector illustrations used to create the figures in this manuscript.

supervised learning. *DNA Res.* 22, 69–77. doi: 10.1093/dnares/dsu041

Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., et al. (2015). Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Syst.* 1, 72–87. doi: 10.1016/j.cels.2015.01.001

Ahmed, S. M., Hall, A. J., Robinson, A. E., Verhoef, L., Premkumar, P., Parashar, U. D., et al. (2014). Global prevalence of norovirus in cases of gastroenteritis: a systematic review and meta-analysis. *Lancet. Infect. Dis.* 14, 725–730. doi: 10.1016/S1473-3099(14)70767-4

- Alavandi, S. V., and Poornima, M. (2012). Viral metagenomics: a tool for virus discovery and diversity in aquaculture. *Indian J. Virol.* 23, 88–98. doi: 10.1007/s13337-012-0075-2
- Amar, C. F. L., East, C. L., Gray, J., Iturriza-Gomara, M., Maclare, E. A., and McLauchlin, J. (2007). Detection by PCR of eight groups of enteric pathogens in 4,627 faecal samples: re-examination of the English case-control infectious intestinal disease study (1993–1996). *Eur. J. Clin. Microbiol. Infect. Dis.* 26, 311–323. doi: 10.1007/s10096-007-0290-8
- Arias, A., Watson, S. J., Asogun, D., Tobin, E. A., Lu, J., Phan, M. V. T., et al. (2016). Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol.* 2:vew016. doi: 10.1093/ve/vew016
- Aw, T. G., Howe, A., and Rose, J. B. (2014). Metagenomic approaches for direct and cell culture evaluation of the virological quality of wastewater. *J. Virol. Methods* 210, 15–21. doi: 10.1016/j.jviromet.2014.09.017
- Beersma, M. F. C., Sukhrie, F. H. A., Bogerman, J., Verhoef, L., Melo, M. M., Vonk, A. G., et al. (2012). Unrecognized norovirus infections in health care institutions and their clinical impact. *J. Clin. Microbiol.* 50, 3040–3045. doi: 10.1128/JCM.00908-12
- Bellou, M., Kokkinos, P., and Vantarakis, A. (2013). Shellfish-borne viral outbreaks: a systematic review. *Food Environ. Virol.* 5, 13–23. doi: 10.1007/s12560-012-9097-6
- Benabbes, L., Ollivier, J., Schaeffer, J., Parnaudeau, S., Rhaissi, H., Nourlil, J., et al. (2013). Norovirus and other human enteric viruses in moroccan shellfish. *Food Environ. Virol.* 5, 35–40. doi: 10.1007/s12560-012-9095-8
- Binnicker, M. J. (2015). Multiplex molecular panels for diagnosis of gastrointestinal infection: performance, result interpretation, and cost-effectiveness. *J. Clin. Microbiol.* 53, 3723–3728. doi: 10.1128/JCM.02103-15
- Boisvert, S., Raymond, F., Godzardis, É., Laviolette, F., Corbeil, J., Wold, B., et al. (2012). Ray meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* 13:R122. doi: 10.1186/gb-2012-13-12-r122
- Briese, T., Kapoor, A., Mishra, N., Jain, K., Kumar, A., Jabado, O. J., et al. (2015). Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *MBio* 6:e1491–15. doi: 10.1128/mBio.01491-15
- Bruni, R., Taffon, S., Equestre, M., Chiionne, P., Madonna, E., Rizzo, C., et al. (2016). Key role of sequencing to trace hepatitis a viruses circulating in Italy during a large multi-country European foodborne outbreak in 2013. *PLoS ONE* 11:e0149642. doi: 10.1371/journal.pone.0149642
- Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176
- Cadwell, K. (2015). The virome in host health and disease. *Immunity* 42, 805–813. doi: 10.1016/j.immuni.2015.05.003
- Cantalupo, P. G., Calgau, B., Zhao, G., Hundesa, A., Wier, A. D., Katz, J. P., et al. (2011). Raw sewage harbors diverse viral populations. *MBio* 2:e180-11. doi: 10.1128/mBio.00180-11
- Colson, P., Richet, H., Desnues, C., Balique, F., Moal, V., Grob, J.-J., et al. (2010). Pepper mild mottle virus, a plant virus associated with specific immune responses, fever, abdominal pains, and pruritus in humans. *PLoS ONE* 5:e10041. doi: 10.1371/journal.pone.0010041
- Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., et al. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.* 5:16532. doi: 10.1038/srep16532
- Cook, N., and Rze, A. (2004). Survival of human enteric viruses in the environment and food. *FEMS Microbiol. Rev.* 28, 441–453. doi: 10.1016/j.femsre.2004.02.001
- Cotten, M., Oude Munnink, B., Canuti, M., Deij, M., Watson, S. J., Kellam, P., et al. (2014). Full genome virus detection in fecal samples using sensitive nucleic acid preparation, deep sequencing, and a novel iterative sequence classification algorithm. *PLoS ONE* 9:e93269. doi: 10.1371/journal.pone.0093269
- D'Agostino, M. D., Cook, N., Rodriguez-Lazaro, D., and Rutjes, S. (2011). Nucleic acid amplification-based methods for detection of enteric viruses: definition of controls and interpretation of results. *Food Environ. Virol.* 3, 55–60.
- de Graaf, M., van Beek, J., and Koopmans, M. P. (2016). Human norovirus transmission and evolution in a changing world. *Nat. Rev. Microbiol.* 14, 421–433. doi: 10.1038/nrmicro.2016.48
- de Wit, M. A., Kortbeek, L. M., Koopmans, M. P., de Jager, C. J., Wannet, W. J., Bartelds, A. I., et al. (2001). A comparison of gastroenteritis in a general practice-based study and a community-based study. *Epidemiol. Infect.* 127, 389–397.
- Deng, X., den Bakker, H. C., and Hendriksen, R. S. (2016). Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu. Rev. Food Sci. Technol.* 7, 353–374. doi: 10.1146/annurev-food-041715-033259
- Di Bartolo, I., Angeloni, G., Ponterio, E., Ostanello, F., and Ruggeri, F. M. (2015). Detection of hepatitis E virus in pork liver sausages. *Int. J. Food Microbiol.* 193, 29–33. doi: 10.1016/j.ijfoodmicro.2014.10.005
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461
- ESFA (2016). Manual for reporting on food–borne outbreaks in accordance with directive 2003/99/EC for information deriving from the year 2015. *EFSA Support Publ.* 13:989E. doi: 10.2903/SP.EFSA.2016.EN-989
- European Food Safety Authority (2014). An update on the risk of transmission of Ebola virus (EBOV) via the food chain. *EFSA J.* 12:3884. doi: 10.2903/j.efsa.2014.3884
- Eurosurveillance editorial team (2015). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2011 has been published. *Euro Surveill.* 18:20449. doi: 10.2903/j.efsa.2015.3991
- Flygare, S., Simon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., et al. (2016). Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol.* 17:111. doi: 10.1186/s13059-016-0969-1
- Frey, K. G., Herrera-Galeano, J., Redden, C. L., Luu, T. V., Servetas, S. L., Mateczun, A. J., et al. (2014). Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC Genomics* 15:96. doi: 10.1186/1471-2164-15-96
- Funk, A. L., Goutard, F. L., Miguel, E., Bourgarel, M., Chevalier, V., Faye, B., et al. (2016). MERS-CoV at the animal–human interface: inputs on exposure pathways from an expert-opinion elicitation. *Front. Vet. Sci.* 3:88. doi: 10.3389/fvets.2016.00088
- Furtak, V., Roivainen, M., Mirochnichenko, O., Zagorodnyaya, T., Laassri, M., Zaidi, S. Z., et al. (2016). Environmental surveillance of viruses by tangential flow filtration and metagenomic reconstruction. *Euro Surveill.* 21:30193. doi: 10.2807/1560-7917.ES.2016.21.15.30193
- Gire, S. K., Goba, A., Andersen, K. G., Sealton, R. S. G., Park, D. J., Kanneh, L., et al. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 345, 1369–1372. doi: 10.1126/science.1259657
- Gossner, C., Gossner, C., and Severi, E. (2014). Three simultaneous, food-borne, multi-country outbreaks of hepatitis A virus infection reported in EPIS-FWD in 2013: What does it mean for the European Union? *Euro Surveill.* 19:20941. doi: 10.2807/1560-7917.ES.2014.19.43.20941
- Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., et al. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.* 7:99. doi: 10.1186/s13073-015-0220-9
- Greninger, A. L., Zerr, D. M., Qin, X., Adler, A. L., Sampoleo, R., Kuypers, J. M., et al. (2016). Rapid metagenomic next-generation sequencing during an investigation of hospital-acquired human parainfluenza virus 3 infections. *J. Clin. Microbiol.* 55, 177–182. doi: 10.1128/JCM.01881-16
- Gruber, K. (2015). Here, there, and everywhere: from PCRs to next-generation sequencing technologies and sequence databases, DNA contaminants creep in from the most unlikely places. *EMBO Rep.* 16, 898–901. doi: 10.15252/embo.201540822
- Guillois, Y., Abravanel, F., Miura, T., Pavio, N., Vaillant, V., Lhomme, S., et al. (2016). High proportion of asymptomatic infections in an outbreak of hepatitis E associated with a spit-roasted piglet, France, 2013. *Clin. Infect. Dis.* 62, 351–357. doi: 10.1093/cid/civ862
- Hald, T., Lo Fo Wong, D. M. A., and Aarestrup, F. M. (2007). The Attribution of human infections with antimicrobial resistant *Salmonella* bacteria in Denmark to sources of animal origin. *Foodborne Pathog. Dis.* 4, 313–326. doi: 10.1089/fpd.2007.0002
- Hamza, I. A., Jurzik, L., Überla, K., and Wilhelm, M. (2011). Methods to detect infectious human enteric viruses in environmental water samples. *Int. J. Hyg. Environ. Health* 214, 424–436. doi: 10.1016/j.ijheh.2011.07.014

- Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685. doi: 10.1128/MMBR.68.4.669-685.2004
- Havelaar, A. H., Kirk, M. D., Torgerson, P. R., Gibb, H. J., Hald, T., Lake, R. J., et al. (2015). World health organization global estimates and regional comparisons of the burden of foodborne disease in 2010. *PLoS Med.* 12:e1001923. doi: 10.1371/journal.pmed.1001923
- Hay, S. I., Battle, K. E., Pigott, D. M., Smith, D. L., Moyes, C. L., Bhatt, S., et al. (2013). Global mapping of infectious disease. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368:20120250. doi: 10.1098/rstb.2012.0250
- Hellmér, M., Paxéus, N., Magnus, L., Enache, L., Arnholt, B., Johansson, A., et al. (2014). Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks. *Appl. Environ. Microbiol.* 80, 6771–6781. doi: 10.1128/AEM.01981-14
- Hingamp, P., Grimsley, N., Acinas, S. G., Clerissi, C., Subirana, L., Poulaïn, J., et al. (2013). Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* 7, 1678–1695. doi: 10.1038/ismej.2013.59
- Hoenen, T., Groseth, A., Rosenke, K., Fischer, R. J., Hoenen, A., Judson, S. D., et al. (2016). Nanopore sequencing as a rapidly deployable ebola outbreak tool. *Emerg. Infect. Dis.* 22, 331–334. doi: 10.3201/eid2202.151796
- Honda, K., and Littman, D. R. (2016). The microbiota in adaptive immune homeostasis and disease. *Nature* 535, 75–84. doi: 10.1038/nature18848
- Iritani, N., Kaida, A., Abe, N., Kubo, H., Sekiguchi, J.-I., Yamamoto, S. P., et al. (2014). Detection and genetic characterization of human enteric viruses in oyster-associated gastroenteritis outbreaks between 2001 and 2012 in Osaka City, Japan. *J. Med. Virol.* 86, 2019–2025. doi: 10.1002/jmv.23883
- Islam, M. S., Sazzad, H. M. S., Satter, S. M., Sultana, S., Hossain, M. J., Hasan, M., et al. (2016). Nipah virus transmission from bats to humans associated with drinking traditional liquor made from date palm sap, Bangladesh, 2011–2014. *Emerg. Infect. Dis.* 22, 664–670. doi: 10.3201/eid2204.151747
- Izopet, J., Dubois, M., Bertagnoli, S., Lhomme, S., Marchandea, S., Boucher, S., et al. (2012). Hepatitis E virus strains in rabbits and evidence of a closely related strain in humans, France. *Emerg. Infect. Dis.* 18, 1274–1281. doi: 10.3201/eid1808.120057
- Karesh, W. B., and Noble, E. (2009). The bushmeat trade: increased opportunities for transmission of zoonotic disease. *Mt. Sinai J. Med.* 76, 429–434. doi: 10.1002/msj.20139
- Kazama, S., Masago, Y., Tohma, K., Souma, N., Imagawa, T., Suzuki, A., et al. (2016). Temporal dynamics of norovirus determined through monitoring of municipal wastewater by pyrosequencing and virological surveillance of gastroenteritis cases. *Water Res.* 92, 244–253. doi: 10.1016/j.watres.2015.10.024
- La Rosa, G., Libera, S. D., Iaconelli, M., Ciccarello, A. R., Bruni, R., Taffon, S., et al. (2014). Surveillance of hepatitis A virus in urban sewages and comparison with cases notified in the course of an outbreak, Italy 2013. *BMC Infect. Dis.* 14:419. doi: 10.1186/1471-2334-14-419
- Lewis, H. C., Wichmann, O., and Duizer, E. (2010). Transmission routes and risk factors for autochthonous hepatitis E virus infection in Europe: a systematic review. *Epidemiol. Infect.* 138, 145–166. doi: 10.1017/S0950268809990847
- Lusk, R. W. (2014). Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS ONE* 9:e110808. doi: 10.1371/journal.pone.0110808
- Mann, E., Streng, S., Bergeron, J., Kircher, A., Morvan, J., Deubel, V., et al. (2015). A review of the role of food and the food system in the transmission and spread of *Ebolavirus*. *PLoS Negl. Trop. Dis.* 9:e0004160. doi: 10.1371/journal.pntd.0004160
- McCloskey, B., Dar, O., Zumla, A., and Heymann, D. L. (2014). Emerging infectious diseases and pandemic potential: status quo and reducing risk of global spread. *Lancet Infect. Dis.* 14, 1001–1010. doi: 10.1016/S1473-3099(14)70846-1
- Menzel, P., Ng, K. L., Krogh, A., Riesenfeld, C., Schloss, P., Handelsman, J., et al. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7:11257. doi: 10.1038/ncomms11257
- Moon, S., Sohn, I.-W., Hong, Y., Lee, H., Park, J.-H., Kwon, G.-Y., et al. (2014). Emerging pathogens and vehicles of food- and water-borne disease outbreaks in Korea, 2007–2012. *Osong Public Health Res. Perspect.* 5, 34–39. doi: 10.1016/j.phrp.2013.12.004
- Newell, D. G., Koopmans, M., Verhoef, L., Duizer, E., Aidara-Kane, A., Sprong, H., et al. (2010). Food-borne diseases – the challenges of 20 years ago still persist while new ones continue to emerge. *Int. J. Food Microbiol.* 139(Suppl.), S3–S15. doi: 10.1016/j.ijfoodmicro.2010.01.021
- Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., et al. (2012). High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *J. Virol.* 86, 12161–12175. doi: 10.1128/JVI.00869-12
- Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2016). metaSPAdes: a new versatile de novo metagenomics assembler. arXiv:1604.03071
- Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 12:385. doi: 10.1186/1471-2105-12-385
- Ott, J. J., Stevens, G. A., Groeger, J., and Wiersma, S. T. (2012). Global epidemiology of hepatitis B virus infection: new estimates of age-specific HBsAg seroprevalence and endemicity. *Vaccine* 30, 2212–2219. doi: 10.1016/j.vaccine.2011.12.116
- Oude Munnink, B. B., Jazaeri Farsani, S. M., Deijls, M., Jonkers, J., Verhoeven, J. T. P., Ieven, M., et al. (2013). Autologous antibody capture to enrich immunogenic viruses for viral discovery. *PLoS ONE* 8:e78454. doi: 10.1371/journal.pone.0078454
- Payne, D. C., Currier, R. L., Staat, M. A., Sahni, L. C., Selvarangan, R., Halasa, N. B., et al. (2015). Epidemiologic Association between FUT2 secretor status and severe rotavirus gastroenteritis in children in the United States. *JAMA Pediatr.* 169, 1040–1045. doi: 10.1001/jamapediatrics.2015.2002
- Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174
- Petrignani, M., Verhoef, L., Vennema, H., van Hunen, R., Baas, D., van Steenbergen, J. E., et al. (2014). Underdiagnosis of foodborne hepatitis A, the Netherlands, 2008–2010. *Emerg. Infect. Dis.* 20, 596–602. doi: 10.3201/eid2004.130753
- Pires, S. M., Evers, E. G., van Pelt, W., Ayers, T., Scallan, E., Angulo, F. J., et al. (2009). Attributing the human disease burden of foodborne infections to specific sources. *Foodborne Pathog. Dis.* 6, 417–424. doi: 10.1089/fpd.2008.0208
- Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* 13:341. doi: 10.1186/1471-2164-13-341
- Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., et al. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol.* 16:114. doi: 10.1186/s13059-015-0677-2
- Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232. doi: 10.1038/nature16996
- Rahman, M. A., Hossain, M. J., Sultana, S., Homaira, N., Khan, S. U., Rahman, M., et al. (2012). Date palm sap linked to nipah virus outbreak in Bangladesh, 2008. *Vector Borne Zoonotic Dis.* 12, 65–72. doi: 10.1089/vbz.2011.0656
- Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F., and Gordon, J. I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* 10, 607–617. doi: 10.1038/nrmicro2853
- Rodriguez-Manzano, J., Hundesa, A., Calgua, B., Carratala, A., Maluquer de Motes, C., Rusinol, M., et al. (2014). Adenovirus and norovirus contaminants in commercially distributed shellfish. *Food Environ. Virol.* 6, 31–41. doi: 10.1007/s12560-013-9133-1
- Sardi, S. I., Somasekar, S., Naccache, S. N., Bandeira, A. C., Tauro, L. B., Campos, G. S., et al. (2016). Coinfections of Zika and Chikungunya viruses in Bahia, Brazil, identified by metagenomic next-generation sequencing. *J. Clin. Microbiol.* 54, 2348–2353. doi: 10.1128/JCM.00877-16
- Schmieder, R., and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 6:e17288. doi: 10.1371/journal.pone.0017288
- Sethi, D., Wheeler, J., Rodrigues, L. C., Fox, S., and Roderick, P. (1999). Investigation of under-ascertainment in epidemiological studies based in general practice. *Int. J. Epidemiol.* 28, 106–112.
- Severi, E., Verhoef, L., Thornton, L., Guzman-Herrador, B. R., Faber, M., Sundqvist, L., et al. (2015). Large and prolonged food-borne multistate hepatitis A outbreak in Europe associated with consumption offrozen berries,

- 2013 to 2014. *Euro Surveill.* 20:21192. doi: 10.2807/1560-7917.ES2015.20.29.21192
- Smits, S. L., Bodewes, R., Ruiz-Gonzalez, A., Baumgärtner, W., Koopmans, M. P., Osterhaus, A. D. M. E., et al. (2015). Recovering full-length viral genomes from metagenomes. *Front. Microbiol.* 6:1069. doi: 10.3389/fmicb.2015.01069
- Spjuth, O., Bongcam-Rudloff, E., Dahlberg, J., Dahlö, M., Kallio, A., Pireddu, L., et al. (2016). Recommendations on e-infrastructures for next-generation sequencing. *Gigascience* 5:26. doi: 10.1186/s13742-016-0132-7
- Tam, C. C., Rodrigues, L. C., Viviani, L., Dodds, J. P., Evans, M. R., Hunter, P. R., et al. (2012). Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut* 61, 69–77. doi: 10.1136/gut.2011.238386
- Tao, Z., Xu, M., Lin, X., Wang, H., Song, L., Wang, S., et al. (2015). Environmental surveillance of genogroup i and ii noroviruses in Shandong Province, China in 2013. *Sci. Rep.* 5:17444. doi: 10.1038/srep17444
- Tavoschi, L., Severi, E., Niskanen, T., Boelaert, F., Rizzi, V., Liebana, E., et al. (2015). Food-borne diseases associated with frozen berries consumption: a historical perspective, European Union, 1983 to 2013. *Euro Surveill.* 20:21193. doi: 10.2807/1560-7917.ES2015.20.29.21193
- Tei, S., Kitajima, N., Takahashi, K., and Mishiro, S. (2003). Zoonotic transmission of hepatitis E virus from deer to human beings. *Lancet* 362, 371–373. doi: 10.1016/S0140-6736(03)14025-1
- Teixeira, J., Mesquita, J. R., Pereira, S. S., Oliveira, R. M. S., Abreu-Silva, J., Rodrigues, A., et al. (2016). Prevalence of hepatitis E virus antibodies in workers occupationally exposed to swine in Portugal. *Med. Microbiol. Immunol.* 206, 77–81. doi: 10.1007/s00430-016-0484-8
- Temmam, S., Davoust, B., Chaber, A.-L., Lignereux, Y., Michelle, C., Monteil-Bouchard, S., et al. (2016). Screening for viral pathogens in african simian bushmeat seized at a French airport. *Transb. Emerg. Dis.* doi: 10.1111/tbed.12481 [Epub ahead of print].
- Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaia, I., Ondov, B., et al. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 14:R2. doi: 10.1186/gb-2013-14-1-r2
- Vázquez-Castellanos, J. F., García-López, R., Pérez-Brocal, V., Pignatelli, M., Moya, A., Handelsman, J., et al. (2014). Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics* 15:37. doi: 10.1186/1471-2164-15-37
- Verhoef, L., Hewitt, J., Barclay, L., Ahmed, S. M., Lake, R., Hall, A. J., et al. (2015). Norovirus genotype profiles associated with foodborne transmission, 1999–2012. *Emerg. Infect. Dis.* 21, 592–599. doi: 10.3201/eid2104.141073
- Verhoef, L., Kouyos, R. D., Vennema, H., Kroneman, A., Siebenga, J., van Pelt, W., et al. (2011). An integrated approach to identifying international foodborne norovirus outbreaks. *Emerg. Infect. Dis.* 17, 412–418. doi: 10.3201/eid1703.100979
- Wang, Y., Zhang, J., and Shen, Z. (2015). The impact of calicivirus mixed infection in an oyster-associated outbreak during a food festival. *J. Clin. Virol.* 73, 55–63. doi: 10.1016/j.jcv.2015.10.004
- Wolfe, N. D., Daszak, P., Kilpatrick, A. M., and Burke, D. S. (2005). Bushmeat hunting, deforestation, and prediction of zoonotic disease. *Emerg. Infect. Dis.* 11, 1822–1827. doi: 10.3201/eid1112.040789
- Wood, D. E., Salzberg, S. L., Venter, C., Remington, K., Heidelberg, J., Halpern, A., et al. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46
- Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053
- Zhang, T., Breitbart, M., Lee, W. H., Run, J.-Q., Wei, C. L., Soh, S. W. L., et al. (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* 4:e3. doi: 10.1371/journal.pbio.0040003
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Copyright © 2017 Nieuwenhuijse and Koopmans. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Characterization of the Genomic Diversity of Norovirus in Linked Patients Using a Metagenomic Deep Sequencing Approach

Neda Nasheri^{1*}, Nicholas Petronella², Jennifer Ronholm^{3,4}, Sabah Bidawid¹ and Nathalie Corneau¹

¹ National Food Virology Reference Centre, Bureau of Microbial Hazards, Food Directorate, Health Canada, Ottawa, ON, Canada, ² Biostatistics and Modeling Division, Bureau of Food Surveillance and Science Integration, Food Directorate, Health Canada, Ottawa, ON, Canada, ³ Department of Food Science and Agricultural Chemistry, Faculty of Agricultural and Environmental Sciences, Macdonald Campus, McGill University, Montreal, QC, Canada, ⁴ Department of Animal Science, Faculty of Agricultural and Environmental Sciences, Macdonald Campus, McGill University, Montreal, QC, Canada

OPEN ACCESS

Edited by:

Abd El-Latif Hesham,
Assiut University, Egypt

Reviewed by:

Hilary G. Morrison,
Marine Biological Laboratory, USA
John W. A. Rosser,
University Medical Center Groningen,
Netherlands

*Correspondence:

Neda Nasheri
neda.nasheri@hc-sc.gc.ca

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 11 October 2016

Accepted: 11 January 2017

Published: 31 January 2017

Citation:

Nasheri N, Petronella N, Ronholm J, Bidawid S and Corneau N (2017) Characterization of the Genomic Diversity of Norovirus in Linked Patients Using a Metagenomic Deep Sequencing Approach. *Front. Microbiol.* 8:73. doi: 10.3389/fmicb.2017.00073

Norovirus (NoV) is the leading cause of gastroenteritis worldwide. A robust cell culture system does not exist for NoV and therefore detailed characterization of outbreak and sporadic strains relies on molecular techniques. In this study, we employed a metagenomic approach that uses non-specific amplification followed by next-generation sequencing to whole genome sequence NoV genomes directly from clinical samples obtained from 8 linked patients. Enough sequencing depth was obtained for each sample to use a *de novo* assembly of near-complete genome sequences. The resultant consensus sequences were then used to identify inter-host nucleotide variations that occur after direct transmission, analyze amino acid variations in the major capsid protein, and provide evidence of recombination events. The analysis of intra-host quasispecies diversity was possible due to high coverage-depth. We also observed a linear relationship between NoV viral load in the clinical sample and the number of sequence reads that could be attributed to NoV. The method demonstrated here has the potential for future use in whole genome sequence analyses of other RNA viruses isolated from clinical, environmental, and food specimens.

Keywords: norovirus, next-generation sequencing (NGS), linked patients, genetic variation, antigenic variation, recombination

INTRODUCTION

Human norovirus (NoV) belongs to *Caliciviridae* family and its genome is comprised of a single stranded, positive sense RNA that contains 3 open reading frames (ORFs) coding for 9 non-structural and structural proteins (Green, 2013). NoV is the most frequent cause of infectious gastroenteritis worldwide (Scallan et al., 2011; Havelaar et al., 2015). While NoV infections are generally acute and self-limiting, persistent infections have been reported in immunocompromised and elderly patients (Aoki et al., 2010; Belliot et al., 2014; Green, 2014).

Abbreviations: NGS, Next-generation Sequencing; NoV, Norovirus; ORF, Open reading frame; RdRp, RNA dependent RNA polymerase; WGS, Whole genome sequencing.

Based on genetic diversity in the capsid region NoVs are divided into 6 genogroups (GI–GVI) which are further divided into 36 genotypes (Kroneman et al., 2011; Vinje, 2015; Sarvestani et al., 2016). Nine genotypes of GI NoVs, 22 genotypes of GII, and a single genotype of GIV and GVI are known to affect humans (Ramirez et al., 2008; Bull and White, 2011; Moore et al., 2015). Genogroup II is the most prevalent genogroup, which accounts for 96% of all infections, and GII.4 is the most common genotype globally (Bull and White, 2011; Eden et al., 2013; Vega et al., 2014). NoVs are known to undergo rapid evolution. The overall mutation rate for the viral capsid protein, which determines the antigenic profile, is 4.16×10^{-3} nucleotide substitutions per site per year and this rate is consistent with other RNA viruses (10^{-3} – 10^{-5} per site per year). The mutation rate is slightly higher for GII.4 (4.3×10^{-3} per site per year), and this is believed to provide the GII.4 strains with higher epidemiological fitness (Bull et al., 2010; Boon et al., 2011). There is evidence that GII.17 is replacing GII.4 as the predominant genotype in certain parts of the world, but detailed epidemiological studies have yet to be conducted to determine the mutation rate of GII.17 (Chan et al., 2015; Chen et al., 2015; de Graaf et al., 2015; Parra and Green, 2015). Genetic diversity also has fitness benefits for NoV, such as increased viral replication, antigenic variability, and enhanced virulence. Antigenic variability is particularly important since antigenic epitopes are under constant evolutionary pressure from the host immune system and novel strain emergence often results in pandemic NoV outbreaks (Karst and Baric, 2015; de Graaf et al., 2016). Therefore, the ability to identify intra- and inter-host NoV genetic diversity and evolutionary hotspots could lead to a better understanding of the processes underlying viral evolution and novel strain emergence, and this has the potential to aid in vaccine development.

NoV can be spread through multiple routes; although, person-to-person and foodborne transmission are the most common (Hall et al., 2013; Barclay et al., 2014; Verhoef et al., 2015). Due to lack of surveillance data, attribution of NoV cases or outbreaks to specific food commodities, as well as determining the direction of NoV transmission after person-to-person spread is often challenging (Barclay et al., 2014; Moore et al., 2015). An effective approach to controlling NoV infection is to understand the mechanism and direction of transmission and then to enact an effective intervention strategy.

Whole genome sequencing (WGS) data can be used to elucidate phylogenetic relationships, monitor the transmission chains and evolution of viruses. For example, during the recent (2014–2016) Ebola outbreak, a WGS approach was adopted to reveal valuable information regarding the viral transmission dynamics, genomic variations (Gire et al., 2014), and allowed for identification of new signatures for prediction of emerging strains (Sozhamannan et al., 2015). Post-outbreak, retrospective WGS analysis can also be performed on NoV strains to examine quasispecies evolution and predict the development of emerging strains through identification of rare variants with predictable signatures that could render predominance. Therefore, WGS analysis of NoV has the potential to provide fast and accurate source attribution, tracking of transmission direction, and

prediction of emerging strains (Kundu et al., 2013; Barclay et al., 2014; Karst and Baric, 2015).

Next-generation sequencing (NGS) approaches now offer routine WGS in research and testing laboratories; however, NGS protocols have yet to be developed for all potential applications (Ronholm et al., 2016). WGS of NoV, directly from clinical samples is routinely accomplished by performing RT-PCR to amplify viral RNA using several sequence-specific and overlapping primer sets (Kundu et al., 2013; Cotten et al., 2014). However, the use of sequence-specific primer sets introduces amplification bias and the assembly of amplicons into WGS assumes conserved viral synteny, this result in overlooking genomic variations. Also, generating whole genome sequences using this approach for even small number of samples is laborious (Thomson et al., 2016). Recently a target enrichment library preparation approach based on SureSelect sequence-specific probes was successfully employed for WGS analysis of NoVs from clinical samples (Brown et al., 2016). While this approach overcomes the problem of primer design in amplicon sequencing, it is costly, requires the prior knowledge of genotype, and is only more effective compared to conventional RNA-Seq approach for samples with low viral loads (Thomson et al., 2016). In this study, we have employed a rapid and unbiased metagenomic NGS approach for NoV from clinical samples, which also provides enough coverage depth for identification of minor variants within viral quasispecies (Mancuso et al., 2011; Iles et al., 2015). Bias was further reduced by using *de novo* assembly. In each case sequencing coverage was sufficient for assembly into a single contiguous sequence. We have also performed a comprehensive genomic analysis during direct transmission of NoV from one patient to another by investigating NoV infection in linked patients. We analyzed the data to provide a detailed description of the viral genome and the presence of inter- and intra-host variants, as well as genome-wide recombination events.

MATERIALS AND METHODS

Sample Collection and Preparation

Fecal samples from families in Ottawa, Ontario who reported symptoms of diarrhea and vomiting were submitted to the National Food Virology Reference Centre at Health Canada for verification of NoV infection. We chose 8 NoV positive samples from 4 families submitted 2013–2015 for our study (Table 1). In all cases the source of original transmission was most likely person-to-person and the direction of transmission was known through the recorded epidemiological data. In family A patient 13–38 infected 13–39, in family B 14–55 infected 14–56, in family C14–58 infected 14–59, and in family D15–65 infected 15–66. In each transmission event the infectious source was a toddler, and the secondary infection was a parent for families A, B, and D, and a sibling for family C.

Ten percent stool suspensions were clarified by centrifugation ($6000 \times g$ for 5 min) and the supernatant was filtered through a $0.45 \mu M$ then $0.22 \mu M$ filter (Millipore, Etobicoke, Ontario, Canada). RNA was extracted from filtrate using the Viral RNA Mini Kit (Qiagen, Mississauga, Ontario, Canada) according to the manufacturer's protocol. Reverse transcription (RT)-PCR was

TABLE 1 | De novo assembly of NoV genomes from clinical samples obtained from linked patients.

Family group	Sample ID	Genotype	Viral Load	Total no. Reads	Total Viral Reads	% Viral Reads	Fold Coverage
A	13–38	GII.4	17640	12,042,724	7163	0.06	976
A	13–39	GII.4	23606	4,301,150	12,373	0.29	121
B	14–55	GII.6	60400	3,950,778	22,326	0.57	223
B	14–56	GII.6	9814	12,493,937	7939	0.06	121.5
C	15–58	GII.4	20159	2,730,201	8766	0.32	75.5
C	15–59	GII.4	61600	3,473,043	29,427	0.85	261
D	15–65	GII.6	6268	5,406,137	2946	0.05	309.5
D	15–66*	GII.6	2706	12,156,220	566	0.00	72

Viral titre is calculated as genome copies/ μ L. *Consensus sequence obtained from 15–65 was used as a reference for the assembly of 15–66.

conducted using a One-Step RT-PCR kit (Qiagen) according to the manufacturer's instructions and primers as described previously (Mattison et al., 2010). Following electrophoresis and in-gel visualization, the amplified products were extracted with the Qiagen gel-purification kit and were sequenced with an ABI3130 Genetic Analyzer to validate NoV infection and for genotype determination.

Viral titres were determined by droplet digital PCR (Bio-Rad, Hercules, California, USA) (Racki et al., 2014) using the probes and primers that were described previously (Kageyama et al., 2003). For this purpose, 20 μ L of each reaction mix was converted to droplets with the QX200 droplet generator (Bio-Rad). Droplet-partitioned samples were then transferred to a 96-well plate, sealed and cycled in a C1000 deep well Thermocycler (Bio-Rad) under the following cycling protocol: 42°C for 30 min (RT) and 95°C for 5 min (DNA polymerase activation), followed by 45 cycles of 95°C for 30 s (denaturation) and 50°C for 1 min (annealing) followed by post-cycling steps of 98°C for 10 min (enzyme inactivation), and then an infinite 4°C hold. The cycled plate was transferred and read in the FAM and HEX channels using the QX200 reader (Bio-Rad).

RNA-Seq Library Preparation

The quality and quantity of extracted RNA was examined using Agilent RNA 6000 Pico Assay Kit and Protocol (Agilent Technologies, Santa Clara, California, USA). Ethanol precipitation of RNA was performed prior to proceeding to TruSeq Stranded mRNA (Illumina, San Diego, California, USA) sample preparation according to the manufacturer's instructions. Briefly, RNA pellets from ethanol precipitates were suspended in Elute, Prime, and Fragment Mix solution, followed by fragmentation at 94°C for 8 min. The first strand cDNA synthesis was conducted using SuperScript III (ThermoFisher Scientific, Waltham, Massachusetts, USA) and random hexamer primers at 25°C for 10 min, 50°C for 50 min, 70°C for 15 min. Second Strand Master Mix was added into the original reaction mixture and used to synthesize the second cDNA strand at 16°C for 60 min. The Agencourt AMPure XP—PCR Purification system (Beckman Coulter Canada, Mississauga, Ontario, Canada) was used to purify the complete double-stranded cDNA. Next A-Tailing Mix was added to the purified cDNA by incubating at 37°C for 30 min to adenylate the 3' ends. Ligation to commercial

adapters, a thymidine overhang from the indexed paired-ends adapters, was conducted at 30°C for 10 min by adding DNA Ligase Mix and adapters to the sample DNA. To stop the ligation, Stop Ligase Buffer was added into the reaction mixture. The Agencourt AMPure XP—PCR Purification system (Beckman Coulter Canada) was used to purify adapter-ligated-DNA products. Library enrichment was performed by adding a PCR MasterMix at 98°C for 30 s and 16 cycles of 98°C for 10 s, 60°C for 30 s, and 72°C for 30 s, with a final extension at 72°C for 5 min. The enriched library was purified by the Agencourt AMPure XP—PCR Purification system (Beckman Coulter Canada), followed by elution with Resuspension Buffer. DNA quantity and quality was measured using Agilent High Sensitivity DNA Assay (Agilent Technologies) as well as the Quant-iT High-Sensitivity dsDNA Assay (ThermoFisher Scientific). The sample preparations were pooled at a concentration of 1 nM of DNA. Freshly diluted NaOH (0.2N) was added to equal volumes of the DNA libraries for denaturation, followed by further dilution with pre-chilled HT1 buffer to obtain a final concentration of 8 pM in a total volume of 1 ml. A 1% final PhiX concentration was added to the denatured DNA for use as an internal control. The prepared DNA library was loaded into a 150 cycle MiSeq Reagent Kit v3 and paired-end sequenced for 76 bp in each direction. The data demonstrated in this article are obtained from the sum of two MiSeq runs. The yield for the first run was 4.52 Gbp (35.1 M reads) and for the second run was 4.78 Gbp (41.3 M reads).

Sequence Assembly and Analysis

The raw sequence reads for each sample consisted primarily of non-NoV reads; therefore, preliminary computational steps needed to be used to ensure an adequate NoV assembly. An in-house database was created that comprised all of the NoV sequences available from NCBI. This database included both complete and partial NoV genomes as well as all available complete and partial open reading frames and coding sequence (CDS) genes. The raw sequence reads for each sample were aligned, using the BLASTn algorithm, to the in-house NoV database, all reads that matched a database entry were binned by the corresponding hit's GenInfo Identifier (GI). Each bin was inspected to identify which NoV database sequences were hit, how many sequence reads were associated with a given GI number, and which complete NoV genome entry had the

most hits from a sample. To produce a *de novo* whole genome assembly from sequence reads matching the NoV database bins were concatenated and concatenated reads were assembled into contigs using SPAdes (Bankevich et al., 2012).

Additionally, the previously identified closed genome with the highest number of associated reads was then used in a reference guided assembly using SMALT v 0.7.4 (<https://sourceforge.net/projects/smalt/>). SMALT was chosen because we have already observed that it performed well with low coverages and when using distant references (Pightling et al., 2014). This was performed to identify variants to the reference genome. Variants were reported using Bcftools (Li, 2011).

The custom script and database used to perform assemblies is available at <https://sourceforge.net/projects/norobin>. The SRA accession numbers of each *de novo* assembly are as follows: SRR3441741, SRR3458065, SRR3458066, SRR3458067, SRR3458068, SRR3458069, SRR3458070, SRR3458071. Throughout this article, we refer to the variations within the viral quasispecies as single nucleotide variations (SNV), and the variations between the consensus sequences obtained from *de novo* assemblies as single nucleotide polymorphisms (SNP).

Phylogenetic Analysis

The NoV consensus sequences obtained from each *de novo* assembly were aligned with chosen reference genomes obtained from GenBank using MUSCLE (Edgar, 2004). The phylogenetic trees were constructed from this alignment with RAxML implementing a GTR Gamma nucleotide substitution model (Stamatakis, 2014) using the sequences from ORF1 and ORF2.

Recombination Analysis

Potential recombination within the complete genome sequences was screened using seven methods (RDP, GENECONV, MaxChi, Bootscan, Chimera, SiScan, and 3Seq) implemented in the Recombination Detection Program version 4.46 (RDP4) (Martin et al., 2015). The breakpoints were also defined by RDP4. Similarity between the recombinants and their possible major and minor parents was estimated using BootScan, embedded in RDP4 (Stamatakis, 2014). SimPlot (Lole et al., 1999) was used to visualize the relationships among the recombinants and their possible parents. The recombination event evaluated by RDP4 was considered significant if it satisfied at least 2 criteria when the $P < 0.05$ and the RDP recombination consensus score (RDPRCS) was $>0.6^{31}$ (Kim et al., 2016).

RESULTS

De novo Assembly of Norovirus Sequences

We obtained 8 fecal samples from 8 symptomatic NoV patients belonging to 4 families (A–D). All samples were RT-PCR positive for NoV GII and further analyses demonstrated that four samples belong to GII.4 genotype (from families A and C) and four samples belong to GII.6 genotype (from families B and D) (Table 1).

Near-full-length genome sequences (78.9–99.9% coverage length) were generated from all 8 samples (GenBank accession

numbers: KX158279, KX158280, KX158281, KX158282, KX158283, KX158284, KX158285, KX158286). The total number of sequence reads from each sample varied between 2.7 and 12.5 million (Table 1) and the proportion of reads that matched a NoV reference sequence also varied between 0.005 and 0.85% of the total sequence reads. This observation is consistent with previous reports for other RNA viruses (Wong et al., 2013). Average coverage depth across genomes was 72- to 976-fold with an average of 270-fold (Table 1). It appears that the minimum number of sequence reads required to obtain an accurate full-length consensus sequence for norovirus was 2900 (Table 1). There was not a correlation between total number of reads and percentage of viral reads ($r = -0.565$). Seven out of 8 sequenced samples produced sufficient reads to allow 90% of the reference genome bases to be called. However, sample 15–66 yielded a low number of reads therefore we used the 15–65 sequence as a reference for the assembly of 15–66.

The Relation between Viral Titre and Depth of Coverage

Whole genome sequence analyses of other RNA viruses have demonstrated that the depth of coverage improves by increasing viral load (Wong et al., 2013; Logan et al., 2014). To elucidate if the correlation between NoV viral titre and the depth of coverage was linear, we plotted the total number of viral reads versus viral titre (genome copies/ μ l). As shown in Figure 1, we observed a strong correlation ($r = 0.972$) between viral titer and depth of coverage. Samples with highest viral load exhibited higher coverage, while samples with lower titre yield low coverage.

Distribution of Viral Reads

We examined genome-wide depth of coverage for each of the genomes by measuring the number of reads per position. The coverage across the length of the genome was not even. Sequence coverage was the highest in the mid ORF1 and ORF2 regions, and lower at the 5' and 3' ends of genomes (Figure 2). This finding is consistent with other studies reporting difficulty in recovering readable fragments from Illumina short-read sequences at the end of DNA molecules (Mortazavi et al., 2008; Batty et al., 2013). The overall coverage profile was consistent between the linked genomes (Figure 2), as well as between genomes from the same genotype (Supplementary Figure 1). This suggests that the depth of coverage can be affected by an intrinsic property of the viral RNA genome. All sequenced genomes had low coverage depth at the intersection of ORF1-ORF2 (Figure 2). This could be explained by the presence of highly conserved and functional structured region at the junction of ORF1-ORF2 (Simmonds et al., 2008; Alhatlani et al., 2015). There is evidence that the 5'-end of ORF2 and the subgenomic RNA (sgRNA) contains extensive RNA secondary structures that function as the promoter for ORF2. Disruption of these evolutionarily conserved RNA stem-loops severely decreases viral replication (Simmonds et al., 2008; Alhatlani et al., 2015). Therefore, our observations further suggest that RNA secondary structures appear to interfere with coverage depth.

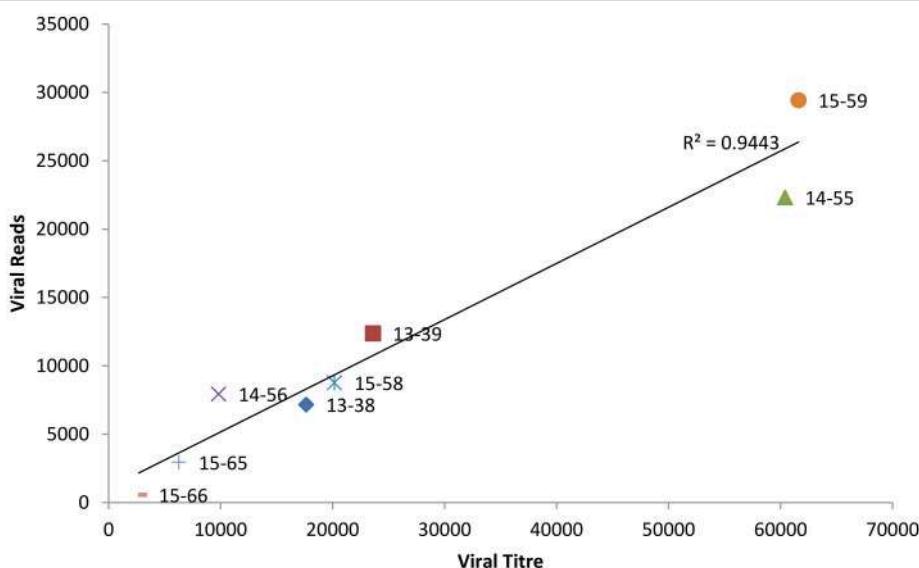


FIGURE 1 | Correlation between the viral titre (genome copies/ μ l) and depth of coverage (reads mapped to norovirus genome).

Phylogenetic Analyses

Phylogenetic trees were constructed from the alignment of the censuses sequences obtained from *de novo* assembly of complete ORF1 (**Figures 3A,B**), and complete ORF 2 (**Figures 3C,D**) with highly similar sequences from the NCBI database. To assess the robustness of each node, a bootstrap resampling process was performed (1000 replicates) and demonstrated at each node. As shown, the phylogenetic relationships between GII.4 sequences (13-38, and 13-39, as well as 15-58, and 15-59, from families A and C respectively) and their closest relatives remain similar for both ORF1 and ORF2 (**Figures 3A,C**). Sequences obtained from 13-38, and 13-39 cluster with GII.4-Hu-CUHK3630-2012-HongKong-CHN (Accession No: KC175323), GII.4-Hu-Sydney2012-Fukuyama-JP-2 (Accession No: KJ196280), and GII.4-Hu-NG1242-2011-JP (Accession No: AB972502). Sequences obtained from 15-58 and 15-59 show homology to GII.4-Hu-Iwate5-2012-JP (Accession No: AB972473) at both ORFs. However, the phylogenetic relationships between GII.6 sequences (14-55, 14-56 from family B and 15-65, 15-66 from family D) and their closest relatives differ slightly between ORF1 and ORF2 (**Figures 3B,D**). It appears that there is more distance between 15-65 and 15-66 sequences and their reference (GII.6-Hu-NHBGR59, Accession No: KU870455) at ORF1 compared to ORF2, as there are 231 SNPs between these genomes and their reference at ORF1, while only 61 SNPs were detected at ORF2 (**Figures 3A,B**). Sequences obtained from 14-55 and 14-56 showed homology to GII.6/2014/Groningen (Accession No: LN854568).

Investigation of SNVs Across the NoV Genome

Multiple lines of evidence suggest that the diversity of the NoV quasispecies in infected individuals with a normal immune system is quite low (Bull et al., 2012; Karst and Baric, 2015).

We observed a high similarity between the consensus sequences obtained from linked patients. However, there were several SNPs between the linked genomes that are listed in **Table 2**. While the consensus sequences obtained from family C (15-58/15-59) showed 100% similarity, the consensus sequences gained from family A (13-38/13-39), family B (14-55/14-56), and family D (15-65/15-66) demonstrated 2, 1, and 4 SNPs across genome, respectively (**Table 2**).

High coverage depth facilitates identifying a variety of unique intra-host variants at different frequencies. The variability of nucleotides observed at each position of the sequenced NoV genomes (with coverage depth of 10-fold or higher) was assessed by analysis of sequence variations at each genome position, and demonstrated in dot plots for each family (**Figure 4**). Since the anticipated sequence error rate for Illumina is approximately 2% (Erik Garrison, 2012; Hasing et al., 2016), genetic variations that occur at frequencies higher than 2% can potentially be natural SNVs. As demonstrated in **Figure 4**, the areas with higher variability are observed at the RdRP, VP1, and the VP2 (ORF3) regions with variant frequency of 5% or higher (**Figure 4**).

When mutations occur in the VP1 (ORF 2) protein, it leads to antigenic variation and subsequent immune evasion (White, 2014). The VP1 protein has three main domains: an N-terminal domain (N), a highly conserved shell domain (S), and a protruding domain (P) which forms surface-exposed spikes on the virus surface. The P domain is further divided into hypervariable domain (P2) and a more conserved P1 domain (P1) (Shanker et al., 2011; Debbink et al., 2012). Within the P2 subdomain 5 highly variable blockade epitopes have been defined (A-E). These epitopes surround the histoblood group antigen (HBGA) binding pocket, Epitope A is considered to be the most important determinant of antigenic variation (Lindesmith et al., 2013). However, other residues adjacent to

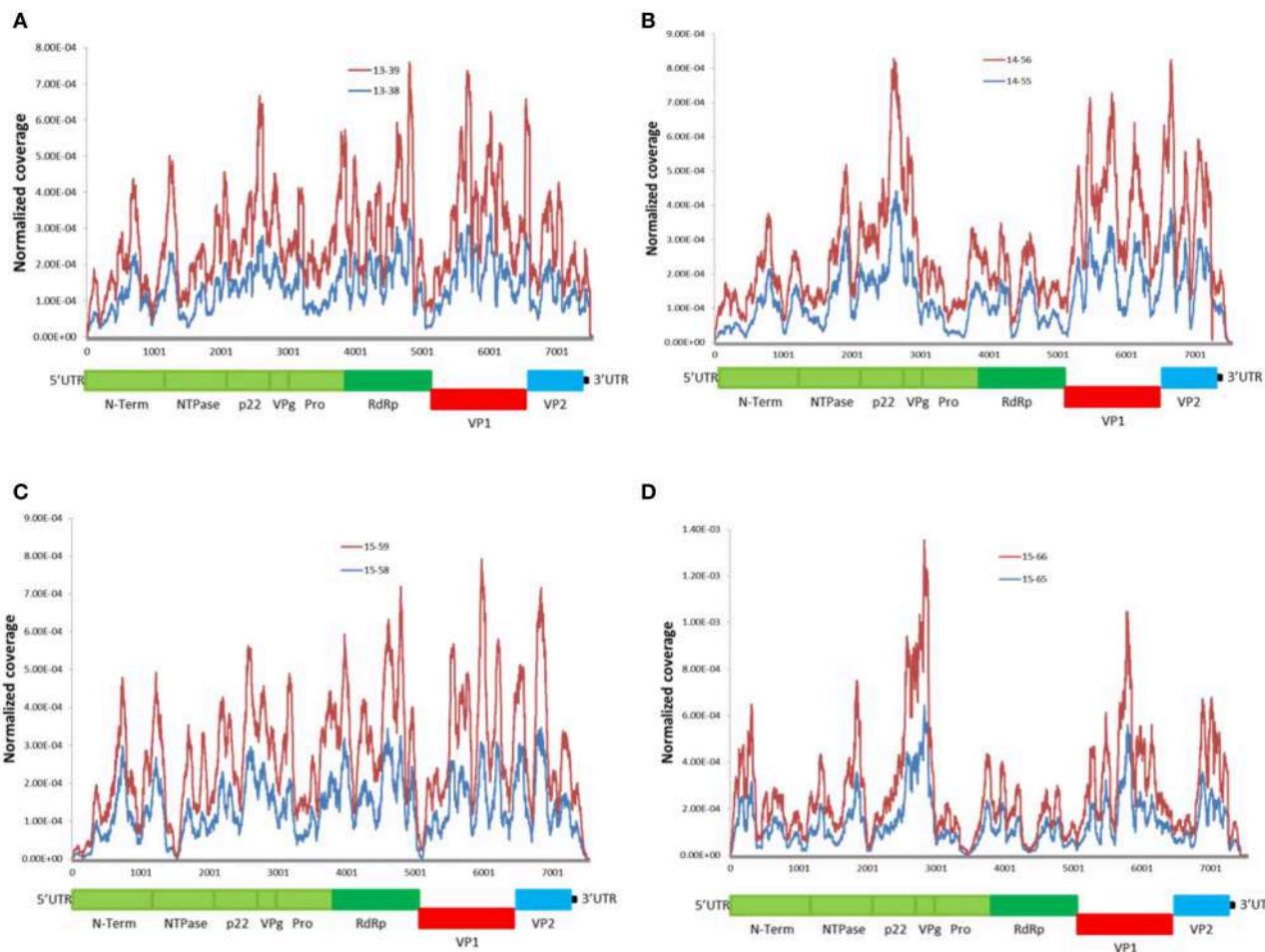


FIGURE 2 | Distribution of NoV reads across the sequenced genomes. Coverage was calculated as the total number of reads covering a given nucleotide and was normalized by the sum of total coverage across the genome, i.e., at each residue, the coverage was divided by the total coverage and the sum of normalized coverage equals one. (A–D) read coverage for the linked patients from families (A–D) (Table 1). Schematic representation of the NoV encoded proteins is shown below the graphs

these epitopes might also impact NoV affinity to HBGAs and binding to neutralizing antibodies (Giammanco et al., 2014; White, 2014).

To study the variation within the VP1 protein between the patients infected with the same genotype, we performed amino acid sequence alignment of VP1 protein (Figures 5A,B). Amino acid sequence alignment of GII.4 Sydney-2012 samples demonstrated no coding change in the conserved N and S domains, 4 in the P1 domain, and 3 in the P2 domain with D376E falling in the C epitope and D391N in the D epitope. Also one non-synonymous change was detected in the C domain (Figure 5A).

Unlike the GII.4 Sydney-2012 isolates, the GII.6 isolates do not belong to the same strain (Figure 3). Therefore, higher amino acid sequence variability is expected in the complete VP1 protein of GII.6 isolates (Figure 5B). The number of non-synonymous differences between samples 14-55, 14-56, and 15-65, 15-66 was 1, 2, 10, 34, and 3 for N, S, P1, P2, and C domains,

respectively. Each blockade epitope had non-synonymous SNPs, with 5 differences in epitope A (T294N, S295A, S297A, V368A, N374D). These differences may have a marked effect on the antigenicity, since it has been previously shown that residues 294 and 368 play key roles in the blockade responses (Vongpunsawad et al., 2013). Two stretches of coding differences that are not within the putative epitopes: one A304P, P305Q, V307A and a relatively longer stretch N351D, T352V, T353S, S355T, S356Q, I357E, G358Q (Figure 5B) were observed. While most of these coding differences exist in the reference genomes, there are several unique coding changes that are only present in the genomes sequenced in this study such as D395I (15-65 and 15-66) and S521P (14-55 and 14-56). Collectively, these data suggest that strains belonging to the same genotype can be antigenically different, and this supports the hypothesis that the emergence of new strains within a certain genotype is the result of escape from herd immunity as they undergo evolution in antigenic and surface-exposing residues.

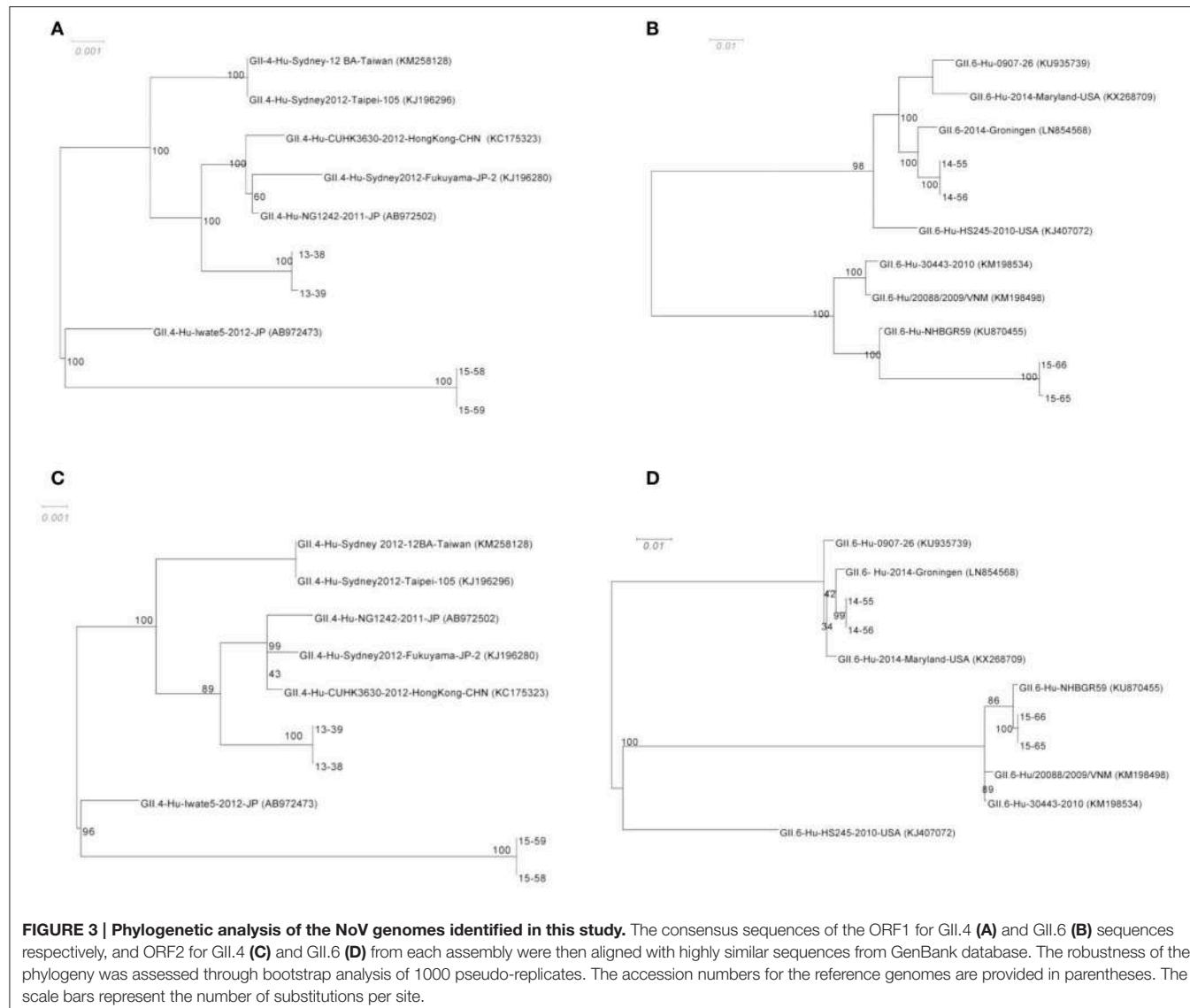


TABLE 2 | Single nucleotide polymorphisms between the linked genomes.

Family	Linked genomes	Position*	Region	Variation
A	13-38/13-39	829	ORF1	G/A
		7447	ORF3	C/T
B	14-55/14-56	3133	ORF1	G/A
C	15-58/15-59	–	–	None
D	15-65/15-66	346	ORF1	G/A
		1102	ORF1	C/T
		1114	ORF1	A/G
		3484	ORF1	C/T

*Nucleotides are numbered according to the NoV GII4/2012/Sydney (Accession No: KM258128).

Recombination Analyses

Recombination is a significant source of genetic diversity in NoVs. This phenomenon, which occurs during co-infection

with different strains, is usually observed at the ORF1-ORF2 and ORF2-ORF3 junctions (White, 2014). There is growing evidence that the recombination rate is increasing among noroviruses (Bull et al., 2012; Wong et al., 2013; Lim et al., 2016); however, the lack of sufficient full-length NoV sequence data has hindered the search for inter- and intra-genotype recombination, and has limited the understanding of how recombination affects NoV evolution (Eden et al., 2013). Therefore, in this study we investigated the incidence of recombination throughout the NoV genomes. Through phylogenetic analysis and separate genotyping of ORF1, and ORF2 regions using Noronet genotyping tool (Kroneman et al., 2011), it was found that all GII.4 Sydney 2012 genomes possessed the original GII.Pe ORF1 (Martella et al., 2013, Table 3) and all GII.6 genomes were associated with GII.P7 ORF1, which is a predominant recombination variant of GII.6 strains (Fumian et al., 2016; Lim et al., 2016).

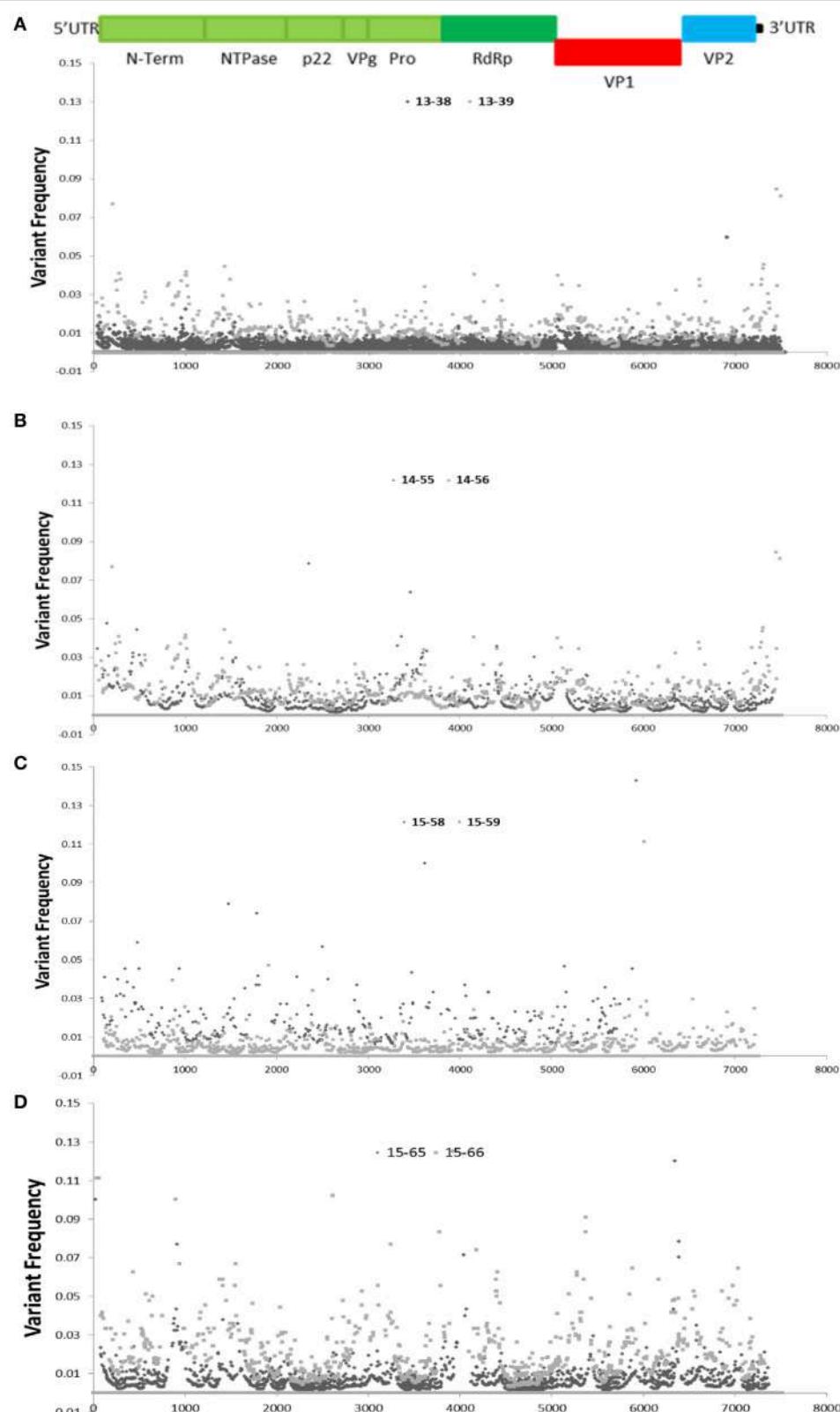
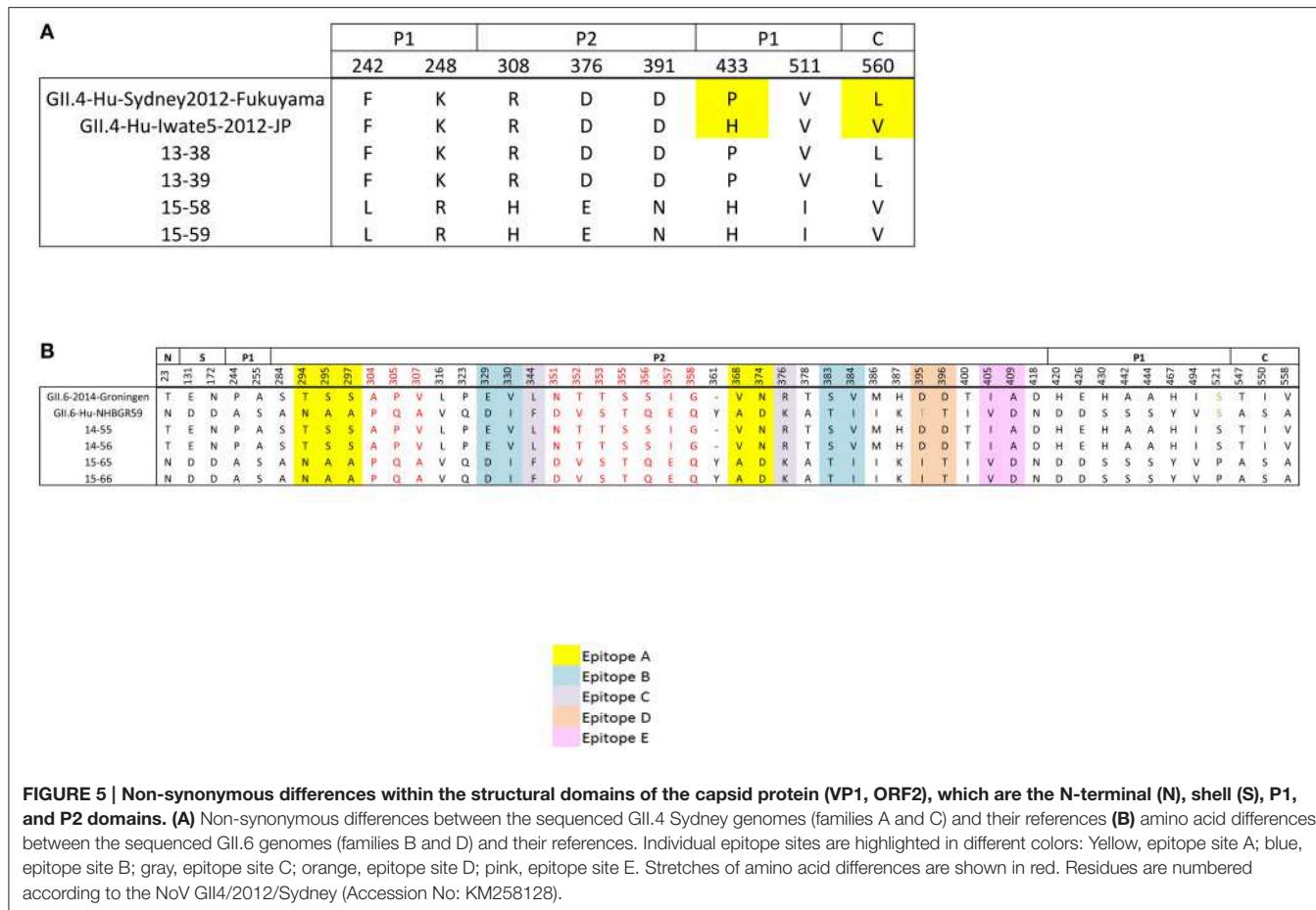


FIGURE 4 | Distribution of single nucleotide variants (SNVs) across the sequenced NoV genomes. Schematic representation of the NoV encoded proteins is shown above the graphs. **(A–D)**, graphs represent SNV frequency at each position for families **(A–D)**.



To perform an in depth analyses of recombination, a genome-wide examination of sequences from the same genotype using RDP4 was preformed (**Figure 6**). This program analyses sequences for the presence of recombination using seven different recombination detection methods (RDP, GeneConv, Bootscan, MaxChi, Chimaera, SiScan, and 3Seq), and provides the statistical probability of each recombination event (Martin et al., 2015). As shown in **Figure 6**, there is evidence of a major recombination event between the 15–65, 15–66, and GII.4 Sydney (Accession No: KM258128) genomes at nucleotide positions 3527–5163, which encompasses the 3' end of the ORF1. This recombination event was unique to 15–65, 15–66 genomes and was not detected in 14–55 and 14–56 genomes. Moreover, further recombination analysis by RDP4 demonstrated that all the GII.4 Sydney genomes (13–38, 13–39, 15–58, and 15–59) contained the same polymerase (GII.Pe), which is considered a signature of the pandemic Sydney 2012 strains (Martella et al., 2013) (data not shown).

DISCUSSION

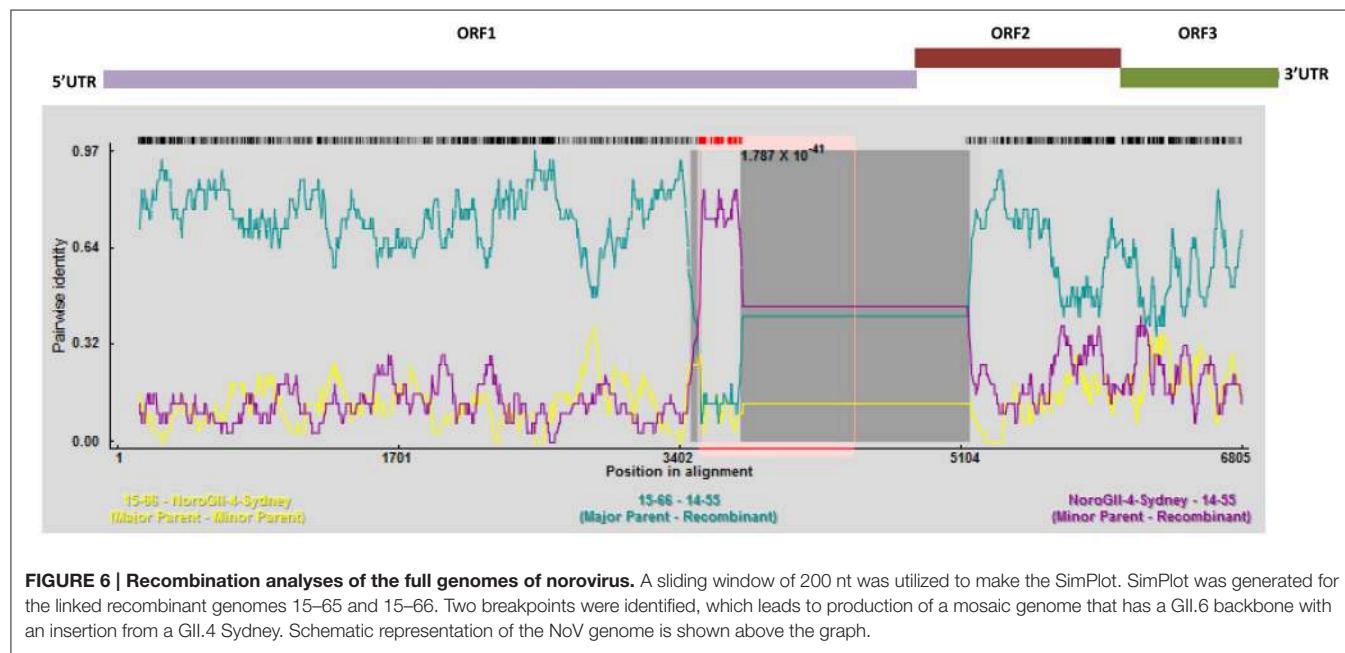
Noroviruses are genetically diverse and this heterogeneity increases the adaptability and fitness of certain variants. Novel antigenic variants can emerge and escape herd immunity

or strains with increased replication rate or environmental persistence can appear that spread more rapidly and efficiently than others. NoVs genetic diversity is influenced by viral factors such as polymerase errors, genetic recombination, and host factors, such as the immune response. It has been hypothesized that inter- and intra-host genetic heterogeneity can act as a reservoir for new variants. Since NoV, like many other RNA viruses, exists as a swarm of quasispecies, it is not clear how much genetic variation can be expected during direct transmission of an acute infection. Therefore, the virus that is isolated from the source of transmission might not necessarily be genetically identical to the virus isolated from the immediate patient. Knowing the range of variation possible between each transmission event is valuable, and our study examined the genetic diversity between a primary and a secondary infection during a single person-to-person transmission events. This is the first step toward facilitating more accurate source attribution as well as establishing a method for determining the directionality of transmission during outbreaks.

Noroviruses generally cause acute and self-limiting infections, therefore we aimed to examine the intra- and inter-patient genetic variations throughout the entire NoV genome isolated from symptomatic, acutely infected, linked patients. For that purpose, next-generation sequencing was used to analyze whole

TABLE 3 | Recombination analysis using Noronet genotyping tool.

Sample ID	Length	ORF1 genotype	ORF2 genotype	ORF1 genotype support	ORF2 genotype support
13-38	7542	GII.Pe	GII.4 Sydney_2012	100	100
13-39	7512	GII.Pe	GII.4 Sydney_2012	100	100
15-58	5963	GII.Pe	GII.4 Sydney_2012	100	100
15-59	7356	GII.Pe	GII.4 Sydney_2012	100	100
15-55	7539	GII.P7	GII.6	100	100
14-56	7491	GII.P7	GII.6	100	100
15-65	7586	GII.P7	GII.6	100	100
15-66	7330	GII.P7	GII.6	100	100



NoV genomes obtained from clinical samples of 8 linked patients (4 GII.4 and 4 GII.6) belonging to 4 families. We selected NoV GII.4 for examination in this investigation because it is the predominant genotype in the North America (de Graaf et al., 2015, 2016), and GII.6 since it is the second most prevalent genotype (Chan-It et al., 2012; Vega et al., 2014; Luo et al., 2015; Xue et al., 2015).

To design a WGS strategy that was unbiased by sequence-specific primers, we employed a metagenomic approach for RNA-Seq on Illumina MiSeq platform to sequence NoV genomes directly from clinical samples. By selecting only sequence-reads that aligned to known NoV sequences, *de novo* assembly of the entire NoV genome was possible (Table 1). High coverage-depth made it possible to identify minor variants at each nucleotide position. We have shown that clinical samples contain sufficient viral RNA to perform successful virus WGS, despite the presence of sequence reads from other sources such as the host, dietary materials, and microbial RNA. Only one sample (15-66), which also had the lowest viral load, did not yield sufficient reads for *de novo* assembly and therefore we performed reference-guided assembly using its linked genome, 15-65, as the reference. Overall

the bases which could not be sequenced were clustered at the ends of the genome, which consistently yields low coverage (Figure 2). Low coverage at the ends of the genome is consistent with other reports that the 5' and 3' terminals of viral RNA genomes are difficult to sequence (Batty et al., 2013). This is particularly evident in samples with low viral titre, suggesting that increasing the input RNA of such samples could improve coverage. We confirmed the correlation between depth of coverage and viral load by plotting coverage against the viral titre and calculating the Pearson correlation coefficient (Figure 1). However, in a study conducted on Measles virus, no correlation was found between the viral titre in samples and the fraction of the genome covered (Penedos et al., 2015).

Variability in coverage and sequencing depth across the genome was observed with all samples; however, overall coverage profile in the linked genomes, as well as genomes that belong to the same genotype, is strikingly similar (Figure 2 and Supplementary Figure 1). While the reason for this observation is not clear, we hypothesized that the presence of RNA secondary structures in the viral genome influences the coverage depth and since the linked genomes have very similar sequences

(**Table 2**), it is assumed that their secondary and tertiary RNA structures are also similar, which yield similar coverage profile. This assumption was further supported by the observation that in all the sequenced genomes, the coverage decreases at the junction of ORF1 and ORF2, where a highly structured promoter for ORF2 exists (**Figure 2**, Simmonds et al., 2008; Alhatlani et al., 2015; Yunus et al., 2015). The negative effect of RNA secondary structure on depth of coverage produced by RNA-Seq has been recently reported by other researchers as well (Biswas and Gao, 2016). RNA structure can affect target accessibility and limit the efficiency of RNA tagmentation and cDNA synthesis. Therefore, a heat-denaturation step prior to RNA tagmentation and cDNA synthesis may improve coverage for WGS of RNA viruses. Moreover, the overall coverage profile obtained from our sample is consistent with what Batty and colleagues have formerly reported (Batty et al., 2013).

For the phylogenetic analysis, a comparison was made in the genetic distance between the genomes sequences assembled in this study and their highly similar relatives for both ORF1 and ORF2 (**Figure 3**). While the genetic distance for the GII.4 Sydney genomes (13–38, 13–39, 15–58, and 15–59) remained consistent for both ORF1 and ORF2, 15–65 and 15–66 sequences that belong to GII.6 appeared to be closer to their reference GII.6-Hu-NHBGR59 in ORF2 when compared to ORF1. This phenomenon can be explained by the presence of a recombination event at the end of ORF1 for these genomes (**Figure 6**).

As expected, there are limited numbers of nucleotide differences between the consensus sequences obtained from linked genomes (**Table 2**). Nevertheless, it is notable that the SNPs that are observed between the linked genomes are missing in the reference genomes (data not shown), this might indicate that these SNPs uniquely occurred during the single transmission events.

Due to high coverage-depth, a variety of unique intra-host SNVs throughout the sequenced genomes was identified. The anticipated sequencing error rate by Illumina technology is approximately 2% (Bravo and Irizarry, 2010; Beerewinkel et al., 2012) therefore it is generally accepted that genetic variations with frequencies $\geq 2\%$ and $\geq 5X$ coverage can be considered as potential SNVs (Erik Garrison, 2012). Other groups have also employed similar approach to identify SNVs within viral quasispecies (Faison et al., 2014; Penedos et al., 2015; Hasing et al., 2016; King et al., 2016). In this study the SNV frequency has been only demonstrated for the positions with 10-fold or higher coverage. While there are low frequency variants that can be considered as sequencing or amplification errors, many of high frequency variants, for example those that clustered within ORF2, the 3' end of ORF1 and ORF3 (**Figure 4**) can potentially be natural occurring intra-patient SNVs.

The distribution of SNVs indicates that intra-host viral populations were more homogenous for patients in families A and B, with only a few SNVs above 5%. In contrast, patients within families C and D presented a more heterogeneous intra-host population, as multiple SNVs with a frequency of 5% or higher in the viral population were distributed across the entire length of the genome. Overall, we observed that some regions of the NoV genomes are particularly prone to mutation

(variant frequency $> 5\%$). It has also been demonstrated that during direct viral transmissions minor variants within the donor viral quasispecies can present as major variants or consensus sequences in the recipient viral population (Kundu et al., 2013; Holzknecht et al., 2015). Therefore, epidemiological data together with deep WGS data obtained from linked cases and transmission chains can provide efficient means for source attribution and may help to identify signatures and patterns in the mutations throughout the viral genomes. In the future this may enable us to predict strain emergence, viral evolution, and allow us to intervene sooner in an outbreak setting.

Amino acid variations in the complete VP1 protein, including the P2 domain were analyzed. Interestingly there are several non-synonymous differences between the closely related GII.4 Sydney genomes (**Figure 5A**), with two amino acid changes falling within the antigenic epitope C and D. These changes can potentially alter the antigenic profile of these viruses and lead to escape from herd immunity. However, in-depth analyses are required to confirm this hypothesis. Amino acid variations were more common between the sequenced GII.6 genomes; specifically the P2 domain contained the most variability with 34 amino acid substitutions (**Figure 5B**). However, the N domain and S domain of the capsid protein were highly conserved. We also screened the antigenic epitope sites for the presence of non-synonymous changes and identified variations in all 5 blockade epitopes, including 5 differences in epitope A, 4 in epitope B, 2 in epitopes C, D, and E. These findings are particularly interesting as these epitopes have been shown to modulate HBGA binding and neutralization responses (Giammanco et al., 2014; White, 2014). Therefore, our data supports the hypothesis that immune evasion by antigenic variation at neutralizing epitopes is a driving mechanism behind new NoV strain emergence.

Most of the recombination events for noroviruses have been observed at the ORF1/2 overlap, which allow the exchange of nonstructural and structural genes, and therefore contribute to the emergence of new epidemic strains (Eden et al., 2013; de Graaf et al., 2016). Herein we also identified GII.P7/GII.6 recombinant viruses that are frequently reported and, unlike the other detected recombinant strains, have long period of circulation (since 2004) (Fumian et al., 2016; Lim et al., 2016). Moreover, we identified a novel inter-genotypic recombination event at the 3' end of ORF 1 for 15–65 and 15–66 genomes (**Figure 6**), which led to the production of a mosaic genome with GII.6 backbone and GII.4 Sydney insertion. This observation is further validated by the phylogenies of the ORF 1 region presented in this study (**Figures 3A,B**). The acquisition of novel ORF1 regions has also been suggested to increase viral fitness that contribute to strain emergence, for example, by altering the replication rate or polymerase fidelity (Eden et al., 2013).

In conclusion, a metagenomic NGS approach can be employed to successfully sequence the whole genome of NoVs directly from clinical samples, without the need for sequence-specific enrichment. In this study, as proof-of-principle, we sequenced each sample to a very high-coverage then recovered and analyzed near-complete genome sequences from 8 linked patients. Processing higher number of samples per MiSeq run, and thus lowering the sequence coverage, would decrease the

sequencing cost. We demonstrated that while *de novo* assembly may not be possible at lower level of coverage, reference guided assembly can be an effective alternative if a close reference exists. Sequence data obtained from this approach can be used to comprehensively analyze intra- and inter-host genetic variation and identify recombination events throughout the NoV genome.

ETHICS STATEMENT

A formal consent was not necessary because the study participants were anonymized.

AVAILABILITY OF SUPPORTING DATA

The data sets supporting the results of this article are included within the article. The Illumina MiSeq short read sequences are deposited in the NCBI database.

AUTHOR CONTRIBUTIONS

NN and JR designed and initiated the project. NN carried out all laboratory works including RNA extraction, quantification and

RNA-Seq. NN and JR prepared the manuscript for publication. NP performed all the bioinformatics analysis. JR, SB, and NC supervised the project. All authors agreed with the final draft of the manuscript. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank Oksana Mykytczuk and Jennifer Harlow for excellent technical support and their assistance in sample preparations. We also thank Dr. Alex Gill and Dr. Erling Rud from the Research Division of the Bureau of Microbial Hazards Health Canada for reviewing the manuscript and offering helpful comments. This work was financially supported by the Research Division of the Bureau of Microbial Hazards, Health Canada. NN acknowledges support from the Visiting Fellow in a Government Laboratory Program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.00073/full#supplementary-material>

REFERENCES

- Alhatlani, B., Vashist, S., and Goodfellow, I. (2015). Functions of the 5' and 3' ends of calicivirus genomes. *Virus Res.* 206, 134–143. doi: 10.1016/j.virusres.2015.02.002
- Aoki, Y., Suto, A., Mizuta, K., Ahiko, T., Osaka, K., and Matsuzaki, Y. (2010). Duration of norovirus excretion and the longitudinal course of viral load in norovirus-infected elderly patients. *J. Hosp. Infect.* 75, 42–46. doi: 10.1016/j.jhin.2009.12.016
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021
- Barclay, L., Park, G. W., Vega, E., Hall, A., Parashar, U., Vinje, J., et al. (2014). Infection control for norovirus. *Clin. Microbiol. Infect.* 20, 731–740. doi: 10.1111/1469-0691.12674
- Batty, E. M., Wong, T. H., Trebes, A., Argoud, K., Attar, M., Buck, D., et al. (2013). A modified RNA-Seq approach for whole genome sequencing of RNA viruses from faecal and blood samples. *PLoS ONE* 8:e66129. doi: 10.1371/journal.pone.0066129
- Beerenwinkel, N., Gunthard, H. F., Roth, V., and Metzner, K. J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.* 3:329. doi: 10.3389/fmicb.2012.00329
- Belliot, G., Lopman, B. A., Ambert-Balay, K., and Pothier, P. (2014). The burden of norovirus gastroenteritis: an important foodborne and healthcare-related infection. *Clin. Microbiol. Infect.* 20, 724–730. doi: 10.1111/1469-0691.12722
- Biswas, A. K., and Gao, J. X. (2016). PR2S2Clust: patched RNA-seq read segments' structure-oriented clustering. *J. Bioinform. Comput. Biol.* 14:1650027. doi: 10.1142/S021972001650027X
- Boon, D., Mahar, J. E., Abente, E. J., Kirkwood, C. D., Purcell, R. H., Kapikian, A. Z., et al. (2011). Comparative evolution of GII.3 and GII.4 norovirus over a 31-year period. *J. Virol.* 85, 8656–8666. doi: 10.1128/JVI.00472-11
- Bravo, H. C., and Irizarry, R. A. (2010). Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics* 66, 665–674. doi: 10.1111/j.1541-0420.2009.01353.x
- Brown, J. R., Roy, S., Ruis, C., Yara Romero, E., Shah, D., Williams, R., et al. (2016). Norovirus whole-genome sequencing by sureselect target enrichment: a robust and sensitive method. *J. Clin. Microbiol.* 54, 2530–2537. doi: 10.1128/JCM.01052-16
- Bull, R. A., Eden, J. S., Luciani, F., McElroy, K., Rawlinson, W. D., and White, P. A. (2012). Contribution of intra- and interhost dynamics to norovirus evolution. *J. Virol.* 86, 3219–3229. doi: 10.1128/JVI.06712-11
- Bull, R. A., Eden, J. S., Rawlinson, W. D., and White, P. A. (2010). Rapid evolution of pandemic noroviruses of the GII.4 lineage. *PLoS Pathog.* 6:e1000831. doi: 10.1371/journal.ppat.1000831
- Bull, R. A., and White, P. A. (2011). Mechanisms of GII.4 norovirus evolution. *Trends Microbiol.* 19, 233–240. doi: 10.1016/j.tim.2011.01.002
- Chan, M. C., Lee, N., Hung, T. N., Kwok, K., Cheung, K., Tin, E. K., et al. (2015). Rapid emergence and predominance of a broadly recognizing and fast-evolving norovirus GII.17 variant in late 2014. *Nat. Commun.* 6:10061. doi: 10.1038/ncomms10061
- Chan-It, W., Thongprachum, A., Khamrin, P., Kobayashi, M., Okitsu, S., Mizuguchi, M., et al. (2012). Emergence of a new norovirus GII.6 variant in Japan, 2008–2009. *J. Med. Virol.* 84, 1089–1096. doi: 10.1002/jmv.23309
- Chen, H., Qian, F., Xu, J., Chan, M., Shen, Z., Zai, S., et al. (2015). A novel norovirus GII.17 lineage contributed to adult gastroenteritis in Shanghai, China, during the winter of 2014–2015. *Emerg. Microbes Infect.* 4:e67. doi: 10.1038/emi.2015.67
- Cotten, M., Petrova, V., Phan, M. V., Rabaa, M. A., Watson, S. J., Ong, S. H., et al. (2014). Deep sequencing of norovirus genomes defines evolutionary patterns in an urban tropical setting. *J. Virol.* 88, 11056–11069. doi: 10.1128/JVI.01333-14
- Debbink, K., Donaldson, E. F., Lindesmith, L. C., and Baric, R. S. (2012). Genetic mapping of a highly variable norovirus GII.4 blockade epitope: potential role in escape from human herd immunity. *J. Virol.* 86, 1214–1226. doi: 10.1128/JVI.06189-11
- de Graaf, M., van Beek, J., and Koopmans, M. P. (2016). Human norovirus transmission and evolution in a changing world. *Nat. Rev. Microbiol.* 14, 421–433. doi: 10.1038/nrmicro.2016.48
- de Graaf, M., van Beek, J., Vennema, H., Podkolzin, A. T., Hewitt, J., Bucardo, F., et al. (2015). Emergence of a novel GII.17 norovirus – End of the GII.4 era? *Euro Surveill.* 20:L21178. doi: 10.2807/1560-7917.ES2015.20.26.21178
- Eden, J. S., Tanaka, M. M., Boni, M. F., Rawlinson, W. D., and White, P. A. (2013). Recombination within the pandemic norovirus GII.4 lineage. *J. Virol.* 87, 6270–6282. doi: 10.1128/JVI.03464-12

- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340
- Erik Garrison, G. M. (2012). Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio GN].
- Faison, W. J., Rostovtsev, A., Castro-Nallar, E., Crandall, K. A., Chumakov, K., Simonyan, V., et al. (2014). Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human genomes. *Genomics* 104, 1–7. doi: 10.1016/j.ygeno.2014.06.001
- Fumian, T. M., da Silva Ribeiro de Andrade, J., Leite, J. P., and Miagostovich, M. P. (2016). Norovirus recombinant strains isolated from gastroenteritis outbreaks in Southern Brazil, 2004–2011. *PLoS ONE* 11:e0145391. doi: 10.1371/journal.pone.0145391
- Giammanco, G. M., De Grazia, S., Terio, V., Lanave, G., Catella, C., Bonura, F., et al. (2014). Analysis of early strains of the norovirus pandemic variant GII.4 Sydney 2012 identifies mutations in adaptive sites of the capsid protein. *Virology* 450–451, 355–358. doi: 10.1016/j.virol.2013.12.007
- Gire, S. K., Goba, A., Andersen, K. G., Sealoff, R. S., Park, D. J., Kanneh, L., et al. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 345, 1369–1372. doi: 10.1126/science.1259657
- Green, K. (2013). “Caliciviridae: the noroviruses,” in *Fields, 6th Edn.*, eds D. M. Knipe and P. Howley (Philadelphia, PA: Lippincott Williams & Wilkins), 949.
- Green, K. Y. (2014). Norovirus infection in immunocompromised hosts. *Clin. Microbiol. Infect.* 20, 717–723. doi: 10.1111/1469-0691.12761
- Hall, A. J., Wikswo, M. E., Manikonda, K., Roberts, V. A., Yoder, J. S., and Gould, L. H. (2013). Acute gastroenteritis surveillance through the National Outbreak Reporting System, United States. *Emerg. Infect. Dis.* 19, 1305–1309. doi: 10.3201/eid1908.130482
- Hasing, M. E., Hazes, B., Lee, B. E., Preiksaitis, J. K., and Pang, X. L. (2016). A next generation sequencing-based method to study the intra-host genetic diversity of norovirus in patients with acute and chronic infection. *BMC Genomics* 17:480. doi: 10.1186/s12864-016-2831-y
- Havelaar, A. H., Kirk, M. D., Torgerson, P. R., Gibb, H. J., Hald, T., Lake, R. J., et al. (2015). World health organization global estimates and regional comparisons of the burden of foodborne disease in 2010. *PLoS Med.* 12:e1001923. doi: 10.1371/journal.pmed.1001923
- Holzknecht, B. J., Franck, K. T., Nielsen, R. T., Bottiger, B., Fischer, T. K., and Fonager, J. (2015). Sequence analysis of the capsid gene during a genotype II.4 dominated norovirus season in one university hospital: identification of possible transmission routes. *PLoS ONE* 10:e0115331. doi: 10.1371/journal.pone.0115331
- Iles, J. C., Njouom, R., Fouopouapouognigni, Y., Bonsall, D., Bowden, R., Trebes, A., et al. (2015). Characterization of hepatitis C virus recombination in cameroon by use of nonspecific next-generation sequencing. *J. Clin. Microbiol.* 53, 3155–3164. doi: 10.1128/JCM.00483-15
- Kageyama, T., Kojima, S., Shinohara, M., Uchida, K., Fukushi, S., Hoshino, F. B., et al. (2003). Broadly reactive and highly sensitive assay for Norwalk-like viruses based on real-time quantitative reverse transcription-PCR. *J. Clin. Microbiol.* 41, 1548–1557. doi: 10.1128/JCM.41.4.1548-1557.2003
- Karst, S. M., and Baric, R. S. (2015). What is the reservoir of emergent human norovirus strains? *J. Virol.* 89, 5756–5759. doi: 10.1128/JVI.03063-14
- Kim, W. K., Kim, J. A., Song, D. H., Lee, D., Kim, Y. C., Lee, S. Y., et al. (2016). Phylogeographic analysis of hemorrhagic fever with renal syndrome patients using multiplex PCR-based next generation sequencing. *Sci. Rep.* 6:26017. doi: 10.1038/srep26017
- King, D. J., Freimanis, G. L., Orton, R. J., Waters, R. A., Haydon, D. T., and King, D. P. (2016). Investigating intra-host and intra-herd sequence diversity of foot-and-mouth disease virus. *Infect. Genet. Evol.* 44, 286–292. doi: 10.1016/j.meegid.2016.07.010
- Kroneman, A., Vennema, H., Deforche, K., v d Avoort, H., Penaranda, S., Oberste, M. S., et al. (2011). An automated genotyping tool for enteroviruses and noroviruses. *J. Clin. Virol.* 51, 121–125. doi: 10.1016/j.jcv.2011.03.006
- Kundu, S., Lockwood, J., Depledge, D. P., Chaudhry, Y., Aston, A., Rao, K., et al. (2013). Next-generation whole genome sequencing identifies the direction of norovirus transmission in linked patients. *Clin. Infect. Dis.* 57, 407–414. doi: 10.1093/cid/cit287
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509
- Lim, K. L., Hewitt, J., Sitabkhan, A., Eden, J. S., Lun, J., Levy, A., et al. (2016). A multi-site study of norovirus molecular epidemiology in Australia and New Zealand, 2013–2014. *PLoS ONE* 11:e0145254. doi: 10.1371/journal.pone.0145254
- Lindesmith, L. C., Costantini, V., Swanstrom, J., Debbink, K., Donaldson, E. F., Vinje, J., et al. (2013). Emergence of a norovirus GII.4 strain correlates with changes in evolving blockade epitopes. *J. Virol.* 87, 2803–2813. doi: 10.1128/JVI.03106-12
- Logan, G., Freimanis, G. L., King, D. J., Valdazo-Gonzalez, B., Bachanek-Bankowska, K., Sanderson, N. D., et al. (2014). A universal protocol to generate consensus level genome sequences for foot-and-mouth disease virus and other positive-sense polyadenylated RNA viruses using the Illumina MiSeq. *BMC Genomics* 15:828. doi: 10.1186/1471-2164-15-828
- Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., et al. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 73, 152–160.
- Luo, L. F., Qiao, K., Wang, X. G., Ding, K. Y., Su, H. L., Li, C. Z., et al. (2015). Acute gastroenteritis outbreak caused by a GII.6 norovirus. *World J. Gastroenterol.* 21, 5295–5302. doi: 10.3748/wjg.v21.i17.5295
- Mancuso, N., Tork, B., Skums, P., Ganova-Raeva, L., Mandoiu, I., and Zelikovsky, A. (2011). Reconstructing viral quasispecies from NGS amplicon reads. *In Silico Biol.* 11, 237–249. doi: 10.3233/ISB-2012-0458
- Martella, V., Medici, M. C., De Grazia, S., Tummolo, F., Calderaro, A., Bonura, F., et al. (2013). Evidence for recombination between pandemic GII.4 norovirus strains New Orleans 2009 and Sydney 2012. *J. Clin. Microbiol.* 51, 3855–3857. doi: 10.1128/JCM.01847-13
- Martin, D. P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015). RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol.* 1:vev003. doi: 10.1093/ve/vev003
- Mattison, K., Sebunya, T. K., Shukla, A., Noliwe, L. N., and Bidawid, S. (2010). Molecular detection and characterization of noroviruses from children in Botswana. *J. Med. Virol.* 82, 321–324. doi: 10.1002/jmv.21682
- Moore, M. D., Goulter, R. M., and Jaykus, L. A. (2015). Human norovirus as a foodborne pathogen: challenges and developments. *Annu. Rev. Food Sci. Technol.* 6, 411–433. doi: 10.1146/annurev-food-022814-015643
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226
- Parra, G. I., and Green, K. Y. (2015). Genome of emerging norovirus, GII.17, United States, 2014. *Emerg. Infect. Dis.* 21, 1477–1479 doi: 10.3201/eid2108.150652
- Penedos, A. R., Myers, R., Hadef, B., Aladin, F., and Brown, K. E. (2015). Assessment of the utility of whole genome sequencing of measles virus in the characterisation of outbreaks. *PLoS ONE* 10:e0143081. doi: 10.1371/journal.pone.0143081
- Pightling, A. W., Petronella, N., and Pagotto, F. (2014). Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* short-read sequence data greatly influences rates of error in SNP analyses. *PLoS ONE* 9:e104579. doi: 10.1371/journal.pone.0104579
- Racki, N., Morisset, D., Gutierrez-Aguirre, I., and Ravnikar, M. (2014). One-step RT-droplet digital PCR: a breakthrough in the quantification of waterborne RNA viruses. *Anal. Bioanal. Chem.* 406, 661–667. doi: 10.1007/s00216-013-7476-y
- Ramirez, S., Giammanco, G. M., De Grazia, S., Colomba, C., Martella, V., and Arista, S. (2008). Genotyping of GII.4 and GIIb norovirus RT-PCR amplicons by RFLP analysis. *J. Virol. Methods* 147, 250–256. doi: 10.1016/j.jviromet.2007.09.005
- Ronholm, J., Nasheri, N., Petronella, N., and Pagotto, F. (2016). Navigating microbiological food safety in the era of whole genome sequencing. *Clin. Microbiol. Rev.* 29, 837–857. doi: 10.1128/CMR.00056-16
- Sarvestani, S. T., Cotton, B., Fritzlar, S., O'Donnell, T. B., and Mackenzie, J. M. (2016). Norovirus infection: replication, manipulation of host, and interaction

- with the host immune response. *J. Interferon Cytokine Res.* 36, 215–225. doi: 10.1089/jir.2015.0124
- Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., et al. (2011). Foodborne illness acquired in the United States—major pathogens. *Emerg. Infect. Dis.* 17, 7–15. doi: 10.3201/eid1701.091101p1
- Shanker, S., Choi, J. M., Sankaran, B., Atmar, R. L., Estes, M. K., and Prasad, B. V. (2011). Structural analysis of histo-blood group antigen binding specificity in a norovirus GII.4 epidemic variant: implications for epochal evolution. *J. Virol.* 85, 8635–8645. doi: 10.1128/JVI.00848-11
- Simmonds, P., Karakasiliotis, I., Bailey, D., Chaudhry, Y., Evans, D. J., and Goodfellow, I. G. (2008). Bioinformatic and functional analysis of RNA secondary structure elements among different genera of human and animal caliciviruses. *Nucleic Acids Res.* 36, 2530–2546. doi: 10.1093/nar/gkn096
- Sozhamannan, S., Holland, M. Y., Hall, A. T., Negron, D. A., Ivancich, M., Koehler, J. W., et al. (2015). Evaluation of signature erosion in ebola virus due to genomic drift and its impact on the performance of diagnostic assays. *Viruses* 7, 3130–3154. doi: 10.3390/v7062763
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033
- Thomson, E., Ip, C. L., Badhan, A., Christiansen, M. T., Adamson, W., Ansari, M. A., et al. (2016). Comparison of next-generation sequencing technologies for comprehensive assessment of full-length hepatitis C viral genomes. *J. Clin. Microbiol.* 54, 2470–2484. doi: 10.1128/JCM.00330-16
- Vega, E., Barclay, L., Gregoricus, N., Shirley, S. H., Lee, D., and Vinje, J. (2014). Genotypic and epidemiologic trends of norovirus outbreaks in the United States, 2009 to 2013. *J. Clin. Microbiol.* 52, 147–155. doi: 10.1128/JCM.02680-13
- Verhoef, L., Hewitt, J., Barclay, L., Ahmed, S. M., Lake, R., Hall, A. J., et al. (2015). Norovirus genotype profiles associated with foodborne transmission, 1999–2012. *Emerg. Infect. Dis.* 21, 592–599. doi: 10.3201/eid2104.141073
- Vinje, J. (2015). Advances in laboratory methods for detection and typing of norovirus. *J. Clin. Microbiol.* 53, 373–381. doi: 10.1128/JCM.01535-14
- Vongpusawad, S., Venkataram Prasad, B. V., and Estes, M. K. (2013). Norwalk virus minor capsid protein VP2 associates within the VP1 shell domain. *J. Virol.* 87, 4818–4825. doi: 10.1128/JVI.03508-12
- White, P. A. (2014). Evolution of norovirus. *Clin. Microbiol. Infect.* 20, 741–745. doi: 10.1111/1469-0691.12746
- Wong, T. H., Dearlove, B. L., Hedge, J., Giess, A. P., Piazza, P., Trebes, A., et al. (2013). Whole genome sequencing and *de novo* assembly identifies Sydney-like variant noroviruses and recombinants during the winter 2012/2013 outbreak in England. *Virol. J.* 10:335. doi: 10.1186/1743-422X-10-335
- Xue, L., Wu, Q., Kou, X., Cai, W., Zhang, J., and Guo, W. (2015). Genome characterization of a GII.6 norovirus strain identified in China. *Infect. Genet. Evol.* 31, 110–117. doi: 10.1016/j.meegid.2015.01.027
- Yunus, M. A., Lin, X., Bailey, D., Karakasiliotis, I., Chaudhry, Y., Vashist, S., et al. (2015). The murine norovirus core subgenomic RNA promoter consists of a stable stem-loop that can direct accurate initiation of RNA synthesis. *J. Virol.* 89, 1218–1229. doi: 10.1128/JVI.02432-14

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Nasheri, Petronella, Ronholm, Bidawid and Corneau. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The Grapevine and Wine Microbiome: Insights from High-Throughput Amplicon Sequencing

Horatio H. Morgan, Maret du Toit and Mathabatha E. Setati *

Department of Viticulture and Oenology, Institute for Wine Biotechnology, Stellenbosch University, Stellenbosch, South Africa

From the time when microbial activity in wine fermentation was first demonstrated, the microbial ecology of the vineyard, grape, and wine has been extensively investigated using culture-based methods. However, the last 2 decades have been characterized by an important change in the approaches used for microbial examination, due to the introduction of DNA-based community fingerprinting methods such as DGGE, SSCP, T-RFLP, and ARISA. These approaches allowed for the exploration of microbial community structures without the need to cultivate, and have been extensively applied to decipher the microbial populations associated with the grapevine as well as the microbial dynamics throughout grape berry ripening and wine fermentation. These techniques are well-established for the rapid more sensitive profiling of microbial communities; however, they often do not provide direct taxonomic information and possess limited ability to detect the presence of rare taxa and taxa with low abundance. Consequently, the past 5 years have seen an upsurge in the application of high-throughput sequencing methods for the in-depth assessment of the grapevine and wine microbiome. Although a relatively new approach in wine sciences, these methods reveal a considerably greater diversity than previously reported, and identified several species that had not yet been reported. The aim of the current review is to highlight the contribution of high-throughput next generation sequencing and metagenomics approaches to vineyard microbial ecology especially unraveling the influence of vineyard management practices on microbial diversity.

OPEN ACCESS

Edited by:

Sandra Torriani,
University of Verona, Italy

Reviewed by:

Vittorio Capozzi,
University of Foggia, Italy

Carmen Portillo,

Universidad Rovira i Virgili, Spain

*Correspondence:

Mathabatha E. Setati
setati@sun.ac.za

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 18 December 2016

Accepted: 21 April 2017

Published: 11 May 2017

Citation:

Morgan HH, du Toit M and Setati ME (2017) The Grapevine and Wine Microbiome: Insights from High-Throughput Amplicon Sequencing. *Front. Microbiol.* 8:820.
doi: 10.3389/fmicb.2017.00820

INTRODUCTION

The conversion of grape juice into wine was first confirmed to be the result of a microbial process by Louis Pasteur in the middle of the nineteenth-century (Barnett, 2003; Jolly et al., 2014; Bokulich et al., 2016b). Since then, the diversity of the vineyard, grape and wine microbiota has been extensively investigated using traditional microbiological methods involving microscopy, cultivation on different agar media and biochemical characteristics. However, the arrival of DNA-based molecular techniques such as polymerase chain reaction (PCR) and the identification of evolutionarily stable molecular marker genes such as ribosomal RNA (rRNA) genes improved our ability to identify microbial species with better resolution and reliability (Justé et al., 2008; Solieri and Giudici, 2008; Cocolin et al., 2013; Sun and Liu, 2014; Wang et al., 2014; Abbasian et al., 2015b).

The bacterial small subunit ribosomal RNA gene (16S rRNA) as well as the fungal ITS1-5.8S rRNA-ITS2 gene have been recognized as the gold standard for estimating the phylogenetic diversity in microbial communities (Justé et al., 2008; Cocolin et al., 2013; Sun and Liu, 2014). Consequently, for the past 3 decades, molecular techniques relying on rRNA genes as target molecules, have been employed in conjunction with culture-dependent methodologies to identify microorganisms after isolation and growth in pure cultures (Esteve-Zarzoso, 1999; Alessandria et al., 2013; Cocolin et al., 2013). To date more than 40 yeast species (Jolly et al., 2014), 50 bacterial species (Barata et al., 2012) and ~70 genera of filamentous fungi (Rousseaux et al., 2014) associated with grapevine and wine fermentation processes have been isolated and identified using traditional culture-based methods. These methods are however extremely laborious, time consuming and often inconsistent and biased (Andorrà et al., 2008; Sun and Liu, 2014). In addition, only species that are able to grow on the culture media and under the cultivation conditions used can be isolated and identified, while species that are in low abundance, those species for which the prevailing cultivation conditions are not conducive, as well as viable but non-culturable (VBNC) cells, are often overlooked (Abbasian et al., 2015b). These limitations in culture-based methods as well as the difference between culturable and *in situ* diversity increased the importance for research into culture-independent molecular approaches (Nocker et al., 2007). Nevertheless, these methods remain important since the microbial species and strains retrieved in such culture-based approaches can be further exploited depending on their biochemical or genetic profiles. Indeed, the wine industry today has access to more than 100 commercial active dry yeast (ADY) strains of *Saccharomyces cerevisiae* that are used as starter cultures for controlled fermentations (Fernández-Espinar et al., 2001; Guzzon et al., 2014). More recently, strains of non-*Saccharomyces* yeasts such as *Torulaspora delbrueckii*, *Metschnikowia pulcherrima*, *Lachancea thermotolerans*, and *Pichia kluyveri*, and several others have been made available as pure starter cultures and in blends with *S. cerevisiae* (Lu et al., 2016; Padilla et al., 2016).

Introduction of PCR-based methods further created opportunities for the development and improvement of several techniques in molecular ecology. The application of molecular techniques allowed researchers to study microbes not on the basis of their ability to grow on certain media types but rather relied on the presence nucleic acids for detection and identification. Such methods, mostly use DNA extracted directly from the environment as a template for PCR, followed by separation and detection for microbial community profiling. Culture-independent methods are often faster, more sensitive and have a higher accuracy than culture-dependent methods (Justé et al., 2008; Lv et al., 2013). These methods include, single-strand conformational polymorphisms (SSCP), denaturing gradient gel electrophoresis (DGGE), terminal restriction fragment length polymorphisms (T-RFLP), and automated ribosomal intergenic spacer analysis (ARISA; Justé et al., 2008; Kovacs et al., 2010; Slabbert et al., 2010; Balázs et al., 2013; Cocolin et al., 2013; Abbasian et al., 2015b). PCR-DGGE was first

applied in wine fermentation by Cocolin et al. (2001) to monitor the diversity and dynamics of yeast populations. Since then, it has remained the most widely used community profiling method in wine fermentation, also including bacteria (Renouf et al., 2006a,b; Cameron et al., 2013). The technique is often employed in combination with culture-dependent methods and has allowed researchers to decipher the complexity and evolution of the microbial population, during berry ripening and throughout the fermentation process (Prakitchaiwattana et al., 2004; Renouf et al., 2005, 2007; Di Maro et al., 2007; Andorrà et al., 2008). Although PCR-DGGE is typically thought to be appropriate for the analysis of less species-rich environments such as grape must, it has low sensitivity (Andorrà et al., 2010) and is unable to detect populations that are present at a relative abundance of <1% of the population (Fasoli et al., 2003; Andorrà et al., 2008). More recently, SSCP (Grube et al., 2011; Schmid et al., 2011; Martins et al., 2014), T-RFLP (Martins et al., 2012; Sun and Liu, 2014), and ARISA (Brežná et al., 2010; Chovanová et al., 2011; Kraková et al., 2012; Pancher et al., 2012; Setati et al., 2012; Ženíšová et al., 2014; Ghosh et al., 2015) have been employed to profile the wine microbial diversity. Culture-independent methods also allow researchers to monitor populations that are numerically under-represented as well as those in the VBNC state (Andorrà et al., 2010; Cocolin et al., 2013). It is critical to monitor such populations as they can influence wine quality. For instance, several studies have demonstrated that strains of *S. cerevisiae*, *Zygosaccharomyces baillii*, and *Brettanomyces bruxellensis* when exposed to SO₂ can enter into VBNC state and survive for more than a month depending on the pH of the environment (Divol and Lonvaud-Funel, 2005; Salma et al., 2013; Capozzi et al., 2016). During this state a spoilage yeast such as *B. bruxellensis* can produce volatile phenols that impart off-odors to the final wine thus rendering it unpalatable (Salma et al., 2013; Capozzi et al., 2016). Although the culture-independent methods have allowed researchers to detect and monitor the evolution of microbial communities, and capture species that were previously not detected, or even misrepresented with culture-dependent methods (Peršoh, 2015), they do have several limitations associated with each of the methods (**Table 1**). Such limitations, e.g., poor band-resolution, co-migration of species, and PCR amplification biases mean that diversity analysis based on these methods still provides a narrow view of the community composition.

Improvements in DNA sequencing expanded the ability of researchers to study the microbial community structure and function with a higher resolution by employing metagenomic approaches. Metagenomics can be defined as the direct genetic analysis of the collective of genomes within an environmental sample (Thomas et al., 2012), this can be achieved either through whole metagenome sequencing or amplicon-based sequencing. Amplicon sequencing, often grouped under the umbrella of metagenomics, is a culture-independent approach for taxonomic, phylogenetic, or functional profiling of microbial communities, accomplished by sequencing specific marker genes amplified directly from environmental DNA without prior enrichment or cultivation of the target population (Franzosa et al., 2015). The innovations in high-throughput, short-amplicon sequencing are

TABLE 1 | A summary of the advantages and disadvantages of PCR-based culture-independent microbial community fingerprinting methods (Arteau et al., 2010; Cocolin et al., 2013).

Methods	Advantages	Disadvantages
Single-strand conformational polymorphisms (SSCP)	<ul style="list-style-type: none"> Distinct bands can be isolated and sequenced No clamped primers and REs required 	<ul style="list-style-type: none"> High rate of re-annealing of single strands with high DNA concentrations
Denaturing gradient gel electrophoresis (DGGE)	<ul style="list-style-type: none"> Ability to target both RNA and DNA 	<ul style="list-style-type: none"> Only intense and well-separated bands can be sequenced
Real-time quantitative PCR (QPCR)	<ul style="list-style-type: none"> Can be applied to RNA and therefore measures viable population 	<ul style="list-style-type: none"> Abundance quantification may be affected by differences in gene expression at different physiological state of the cells Requires species specific primers
Terminal restriction fragment length polymorphisms (T-RFLP)	<ul style="list-style-type: none"> Easily applicable to large sample numbers Web-based tools allow <i>in silico</i> prediction of TRFs 	<ul style="list-style-type: none"> Incomplete and non-specific digestion leads to overestimation of diversity Poor resolution of complex communities Requires multiple RE's
Automated ribosomal intergenic spacer analysis (ARISA)	<ul style="list-style-type: none"> Less labor intensive Allows detection of dominant species Allows high resolution of subtle differences 	<ul style="list-style-type: none"> Co-migration of species with same ITS amplicon size Preferential amplification of shorter templates

revolutionary in a way that they can describe the microbial diversity within and across complex biomes (Bokulich et al., 2013b). Although high-throughput sequencing technologies have been widely used to investigate the microbial ecology of various environments (Ma et al., 2015; Shi et al., 2015; Abbasian et al., 2015a), their application in grapevine and wine fermentation microbial ecology is relatively recent, and their contribution to the field has not been explored. In recent studies it was also shown that grape microbial diversity is driven by cultivar, climatic conditions both macro- and micro-climate, the seasonal environmental conditions, viticultural farming practices as well as wine microbiome by fermentation process applied during the winemaking (Bokulich et al., 2014; David et al., 2014; Gilbert et al., 2014; Setati et al., 2015; Zarraonaindia et al., 2015; Abbasian et al., 2015a; De Filippis et al., 2017). Therefore, with this review, we aim to provide an in-depth overview of the vineyard, grape, and wine microbiome and its functional potential as unraveled through high-throughput sequencing techniques.

NEXT-GENERATION SEQUENCING

For many years, microbial community analyses relied on the isolation and identification of individual species, or cloning and sequencing of rRNA genes retrieved by PCR from environmental DNA. These methods mainly relied on first-generation DNA sequencing technology which was developed by Sanger et al. (1977). A few decades later, deep high-throughput, in-parallel sequencing technologies collectively referred to as Next-generation sequencing (NGS) were developed (Bleidorn, 2015). The term NGS therefore specifically refers to non-Sanger-based second and third generation sequencing (TGS) techniques (Türktaş et al., 2015).

After Sanger introduced the chain-terminator DNA sequencing method, commercial second generation sequencing (SGS) platforms were developed. The Genome Sequencer 20 system launched in 2005 by 454 Life Sciences, was the first commercial SGS platform and was soon followed by the Genome Analyzer II launched by Solexa/Illumina in 2006. Both these platforms use a sequencing by synthesis approach. Roughly 2 years later, Lifetechnologies/Applied Biosystems introduced the

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) platform which applies fluorophore labeled oligonucleotide panels and ligation chemistry for sequencing. Subsequently, Complete Genomics developed the CGA sequencing technology which employed semi-ordered array of “DNA nanoballs” on a solid surface, while the Ion Torrent, which is regarded as the first of the “post-light sequencing” technologies, was introduced in 2010 (Reuter et al., 2015; Heather and Chain, 2016). The Ion Torrent’s semiconductor sequencer is thought to be a technology between second and TGS categories. The technology is capable of sequencing single molecules thus negating the requirement for prior DNA amplification (Heather and Chain, 2016).

The majority of SGS technologies however, still have various limitations, such as errors arising from PCR (Peršoh, 2015), the loss of synchronicity “dephasing” (Schadt et al., 2010; Diaz-Sanchez et al., 2013) and the duration of completion “time to results” (Diaz-Sanchez et al., 2013). To overcome these drawbacks TGS or next-next generation platforms such as Single-molecule real-time (SMRT) sequencing (Schadt et al., 2010; Bleidorn, 2015) and Nanopore DNA Sequencer (Diaz-Sanchez et al., 2013), which open the possibility for single molecule sequencing were developed. These come with several advantages, (i) higher throughput, (ii) faster “time-to-result,” (iii) low cost, (iv) longer read length, (v) increased consensus accuracy enabling rare variant detection and (vi) small starting material (Schadt et al., 2010; Diaz-Sanchez et al., 2013; Bleidorn, 2015). However, these sequencing methodologies are still in development, and/or in the beta stage. Few commercial platforms have been evaluated, however they remain plagued by high error rates, and low output, although the technology is promising (Bleidorn, 2015). As such they cannot yet replace SGS, which remain and continue to be pivotal in microbial ecology surveys.

NEXT-GENERATION SEQUENCING IN MICROBIAL ECOLOGY

SGS platforms have revolutionized the landscape of microbial ecology and have been the cornerstone of many phylogenetic surveys. The methods make it possible to compare and analyze the whole microbial community diversity, abundance,

and functional genes at far greater sequencing depths. These technologies depend on a parallel process in which each single DNA fragment is sequenced independently and separated in clonal amplicons for downstream analysis between the total sequences generated (Wooley et al., 2010; Diaz-Sanchez et al., 2013). With most SGS methodologies, an uninterrupted operation of a washing and scanning process is used to read tens of thousands of matching strands that are fixed to a specific location (Schadt et al., 2010). The length of the fragments obtained from the analyses differs depending on the sequencing method employed (Wooley et al., 2010; Bokulich et al., 2016b). Until recently, the Illumina and 454 pyrosequencing platforms were the most commonly used platforms for grapevine ecology surveys. At least 48% of the published data on the vineyard, grapevine and wine microbiome is derived from 454 pyrosequencing while the remaining 52% is derived from Illumina sequencing. Both platforms work on a sequencing-by-synthesis approach, however differ in their chemistries.

ILLUMINA

The process of Illumina sequencing, consists of the bridge amplification of adapter-ligated DNA fragments on the surface of a glass (Pettersson et al., 2009). Bases are then determined using a cyclic reversible termination technique, which sequences the template strand, a single nucleotide at a time through progressive rounds of base incorporation, washing, scanning, and cleaning. In this method, labeled dNTPs are used to stop the polymerization reaction, allowing the removal of unincorporated bases. The fluorescent dye is captured to identify the bases added, and then cleaved so that the next nucleotide can be added, this is then repeated (Pettersson et al., 2009; Diaz-Sanchez et al., 2013; Reuter et al., 2015; Heather and Chain, 2016). The earlier Illumina analyser generated at least 1 Gb of sequences with reads averaging 35 bp and the duration of 2–3 days. However, the introduction of HiSeq and MiSeq machines altered the duration time to ~4 days and 24–30 h, and increased the read length to 250–300 bp, respectively with error rates of below 1%, with substitution the most occurring issue (Bleidorn, 2015; Goodwin et al., 2016).

PYROSEQUENCING

In 454 pyrosequencing an emulsion PCR is used for bridge amplification of adapter-ligated DNA fragments on the surface of a bead. The beads are thereafter distributed and fixed into 44 μm wells, where the sequencing by synthesis occurs. After the nucleotide bases are incorporated an enzymatic luciferase coupled reaction occurs, allowing for the identification of bases, which is measured using a charged couple device (Pettersson et al., 2009; Diaz-Sanchez et al., 2013; Reuter et al., 2015; Heather and Chain, 2016). The Roche 454 FLX platform has the ability to generate 80–120 Mb of sequences averaging in 200–300 bp reads, for a run that averages ~4 h with an error rate of below 1% (Morozova and Marra, 2008), while the FLX titanium is capable of producing read lengths of over 400 bp (Pettersson et al., 2009).

The 454 pyrosequencing technique was reported in 2008, as the most published NGS platform, however, the technology has

since been discontinued, and has therefore been surpassed by Illumina which is currently considered to have made the largest contribution to SGS (Huse et al., 2007; Morozova and Marra, 2008; Reuter et al., 2015; Heather and Chain, 2016).

APPLICATION OF NEXT-GENERATION SEQUENCING IN DECIPHERING THE VINEYARD MICROBIOME

The vineyard microbiome broadly describes the collective genomes of microorganisms present in the vineyard ecosystem, including those present in soil, grapevine, cover crops, and the insects associated with the plants. Furthermore, microbial transfer from nearby plants, could be transported aerially or via insects (Gilbert et al., 2014). Consequently, the grape microbiome represents a reservoir of microorganisms comprising filamentous fungi, yeast as well as bacteria. These populations are however variable and are influenced by various external factors, such as grape cultivar, climatic conditions, farming practices, and the vineyard location (Setati et al., 2012; Salvetti et al., 2016). The past decade has seen a significant advancement in the manner in which researchers understand the microbial ecology of the vineyard, due to molecular profiling techniques that have further evolved, to explore microbial ecosystems (Bokulich et al., 2012). Recent studies have employed SGS to decipher the grape and grapevine associated microbiome (David et al., 2014; Pinto et al., 2014), and to determine how viticultural practices could potentially influence these communities (Setati et al., 2015; Kecskeméti et al., 2016; Marzano et al., 2016), their dynamics throughout grape berry development and wine fermentation (Piao et al., 2015; Stefanini et al., 2016) and to unravel their functional potential (Salvetti et al., 2016).

For the comprehensive evaluation of the vineyard and the grape microbiome, two key questions are typically addressed. Firstly, which microorganisms are present within the environment, and secondly the role of the individual species (Ravin et al., 2015). To understand what role the identified species, if any; plays in the grape and wine microbiome requires that standard microbiological methods be applied to isolate the strains and then evaluate them for their potential contribution to grape or wine quality by assessing their phenotypic and genotypic properties and thereafter they will be evaluated in different wine matrices to assess their growth and metabolic profile. To this effect, several species retrieved using culture-dependent methods have been shown to contribute positively in the winemaking process. For instance, some strains of *Wickerhamomyces anomalus*, *Candida pyralidae*, *T. delbrueckii*, and *Kluyveromyces wickerhamii* were shown to suppress the growth of *B. bruxellensis* (Comitini et al., 2017), a wine spoilage yeast; *M. pulcherrima* was highlighted as a desirable co-inoculant for the reduction of ethanol (Morales et al., 2015), while others such as *Hanseniaspora vineae*, *Starmerella bacillaris*, *L. thermotolerans*, *P. kluveri*, and *T. delbrueckii* present various desirable aroma signatures (Jolly et al., 2014; Comitini et al., 2017). In order to explore the untapped diversity revealed by SGS, it would be important to establish enrichment methods that can allow retrieval of those species that have not yet been characterized. Consequently,

different sampling strategies are employed depending on what question the researcher seek to address.

SAMPLING STRATEGIES

The vineyard and grapevine microbiome has been studied from a variety of samples including the soil and different parts of the vines. However, there is currently no standardized sampling strategy or experimental design for vineyard microbiome analysis. For the soil microbiome samples are typically derived from surface soil or from the root zone. Typically, anything from 3 to 5 samples are randomly collected, sifted through a 0–2 mm sieve and then homogenized and composited. Samples are often collected with a spade or with the aid of a 33 inch by 7–8-inch corer, within the alleyways of the vineyard or at a distance of 15–30 cm away from the trunk, at a depth of 0–7 cm (Martins et al., 2014; Burns et al., 2015; Zarraonaindia et al., 2015). In contrast, root soil samples are collected closer to the stem (10–15 cm) although at similar depth to the surface samples (Zarraonaindia et al., 2015). For microbial evaluation of plant material such as roots and branches (Campisano et al., 2014), grapevines of similar age and size are typically chosen, eliminating one source of microbial variability. Only a certain area of the vine is sampled, the material typically peeled or crushed under aseptic conditions for further evaluation. For instance, some studies have used leaves (Leveau and Tech, 2011; Pinto et al., 2014) while others have used the cane, graft union of the trunk as well as the roots (Faist et al., 2016), depending on the aim of the study. In contrast, sampling for analysis of the grape-associated microbiome can vary from a few bunches to kilograms of grapes (David et al., 2014; Taylor et al., 2014; Pinto et al., 2015; Setati et al., 2015; Wang et al., 2015; Salvetti et al., 2016). Careful selection of healthy and undamaged grapes is often critical unless the aim is to investigate botrytized wines (Bokulich et al., 2012) and/or sweet wines (Stefanini et al., 2016). The grapes are subsequently crushed under aseptic conditions and the DNA extracted from the resulting must. In a few cases, samples were collected from commercial wineries as composite grape must (Bokulich et al., 2014, 2016a). In a few studies that monitored population dynamics during fermentation, additional samples are withdrawn at various time points representing the beginning, middle, and end of fermentation (David et al., 2014; Pinto et al., 2015; Wang et al., 2015). In most instances, sample volumes ranging from 5 to 50 mL are then further used for DNA extractions.

TARGET GENES

The target marker genes are universally present in all species evaluated and contain both highly conserved fragments that facilitate the design of PCR primers targeting all members of a community and variable regions that allow for the discrimination of different species within the community (Justé et al., 2008; Cocolin et al., 2013; Sun and Liu, 2014; Wang et al., 2014). In both fungi and bacteria, ribosomal RNA genes are suitable target genes. In bacteria, the 16S rRNA is typically targeted while in fungi the ITS1-5.8S rRNA-ITS2 as well as the 26S rRNA are the

target molecules for high throughput amplicon sequencing and microbiome analyses.

The 9 hypervariable regions (V1–V9) of bacteria have all been targeted for the estimation of vineyard bacterial diversity (Leveau and Tech, 2011; Campisano et al., 2014; Perazzolli et al., 2014; Bokulich et al., 2015, 2016a; Burns et al., 2015; Calleja-Cervantes et al., 2015; Piao et al., 2015; Pinto et al., 2015; Zarraonaindia et al., 2015; Holland et al., 2016; Marzano et al., 2016; Portillo et al., 2016). Depending on the region sequenced the data might be similar or differ significantly. For instance, in a study comparing the V4 and V5 region Bokulich et al. (2012), found that the regions resulted in a similar bacterial composition with minor variation in the lower taxa; although the V4 region provided greater taxonomic depth for certain *Proteobacteria* and lactic acid bacteria (LAB) species. In contrast, Campanaro et al. (2014), targeted the V3–V4 and V5–V6 regions of the 16S rRNA region and evaluated the bacterial community associated with grape marc after crushing and 30 days “post fermentation”/storage. A total of 89 genera were identified, however only 31 of these were common in both target regions evaluated.

The fungal ITS regions are the most commonly targeted region for fungal diversity estimation. The classification of general fungi and arbuscula mycorrhizae (AMF) has been accomplished by targeting the ITS region (Bokulich et al., 2013a, 2015, 2016a; Setati et al., 2015; Bouffaud et al., 2016; Holland et al., 2016; Kecskeméti et al., 2016; Marzano et al., 2016; Stefanini et al., 2016), D1–D2 regions of the 26S rRNA (Holland et al., 2014; Taylor et al., 2014) and the partial 18S rRNA gene (Lumini et al., 2010; David et al., 2014; Holland et al., 2016; Grangeau et al., 2017; De Filippis et al., 2017). The AMF populations derived from these different targets, were similar in genera and showed compositional differences in samples evaluated, highlighting them all as suitable target genes for AMF evaluation (Lumini et al., 2010; Bouffaud et al., 2016). Furthermore, Pinto et al. (2014, 2015) targeted both the ITS2 region and D2 domain of the 26S rRNA region for fungal community analysis. The results revealed that the taxonomic depth for the two evaluated regions was considerably similar, however of these only a portion of the observed OTU's were shared between the two regions and that overall the ITS region provided a slightly higher coverage. Bokulich and Mills (2013) moreover, evaluated several ITS primers, and they found that targeting the ITS1 region demonstrates higher levels of taxonomic classification accuracy (species and genus), the smallest difference between Ascomycota and Basidiomycota amplicon lengths, as well as a maximized sequence coverage. Therefore, overall the ITS1 locus appears to be the most promising target, for a complete overview of the microbial populations in ecological studies.

BIOINFORMATICS AND ANALYSIS

High throughput sequencing techniques generally generate large amounts of sequence data, and the only viable option to handle such information, is via automated approaches. There are currently several open source pipelines accessible for overseeing, almost the complete analysis procedure for NGS data. These include MOTHUR, quantitative insights into microbial

ecology (QIIME; Köljalg et al., 2013), metagenomics rapid annotation using subsystem technology (MG-RAST), server and rapid analysis of multiple metagenomes with clustering and annotation pipeline (RAMMCAP; Wooley et al., 2010). These pipelines provide the tools for basic data analysis steps such as data cleaning, sequence clustering, functional annotation, and taxonomic assignments (Köljalg et al., 2013).

The current section will provide brief overview in the procedures used to analyze high-throughput sequencing data in targeted amplicon sequencing for the vineyard and wine associated microbiome, followed by a brief overview of whole-metagenomics sequencing.

TARGET/AMPLICON SEQUENCING

The analysis of amplicon sequencing data typically undergoes three basic steps; (i) Quality trimming and de-noising; (ii) OTU-picking/clustering, and (iii) taxonomic assignment. Quality-trimming is an essential step used to eradicate erroneous reads obtained through PCR, sequencing instruments and the chemistries behind the sequencing reactions (Bokulich et al., 2013a). To minimize the volume of data for annotation, clustering, and OTU-picking is used. During clustering, pairwise comparison of sequencing is performed with a set percentage identity threshold. Subsequently, a single representative of highly similar sequences is chosen and annotated through BLAST or BLAT algorithms. OTUs can be processed through an open-reference or closed-reference OUT-picking approach. Assignment of species or annotation of functional genes is based on percentage similarity to sequences in specific databases such as Greengenes, UNITE, SILVA, NCBI, SWISSPROT etc.

The analysis of data derived from pyrosequencing during quality trimming typically involves; the removal of barcodes, adapters, and primers, followed by denoising which is used to correct problems associated specifically with 454 pyrosequencer. These typically include the removal of sequences, with ≥ 6 homopolymers, ambiguous bases and those not meeting Phred score of (20–30). Furthermore, sequences of min and max length can be removed, depending on the target region and possible chimeric sequences (**Figure 1**).

The data derived from Illumina sequencing platforms undergoes similar demultiplexing and quality trimming apart from denoising. Reads are typically truncated for ≥ 3 consecutive bases with a quality $<1e^{-5}$, and removed when containing ambiguous base calls, primer/barcode errors or a phred score of <20 –30. Furthermore, for paired-end sequencing, the reads are typically joined after quality trimming prior to OTU picking, with all sequences retained, even those not overlapping (**Figure 1**).

SHOTGUN METAGENOMICS SEQUENCING

While the goal in the analysis of the metagenomic data is to reconstruct all the genomes within the environmental sample, the computational intricacy involved makes it unfeasible.

Thus, as an alternative two general types of analyses are performed for reconstruction; (i) assembling the reads into contigs, and performing taxonomic classification and functional assignments; (ii) read-based reconstruction of the taxonomic and functional parts of the metagenome. During the assembly of sequences, several problems could arise; for instance, limitation in computational space (Peršoh, 2015), formation of chimeras as a consequence of similarities amongst genomes of related species and variable abundances of genomes within the sample which could potentially result in partial representation (Scholz et al., 2012; Ravin et al., 2015).

Since a mixture of varying amounts of genomic fragments, from different organisms is the result of contig assembly, taxonomic classification can be complicated. Nevertheless, clustering based on the nucleotide composition and coverage carried out by different techniques could sort/bin metagenomic data based on taxonomic status. The clustering efficacy does however rely on various factors. Furthermore, the taxonomic status of the resulting “bins” of contigs is obtained through the identification of phylogenetic marker genes in the bin which was analyzed (Ravin et al., 2015). Additional algorithms have been proposed as an alternative to the cluster based algorithms (Kriseman et al., 2010).

The annotation of the metagenomic contigs can be done using various command-line pipelines and online annotations services, such as MG-RAST, integrated microbial genomes and microbiomes (IMG-M) and community cyberinfrastructure for advanced microbial ecology research and analysis (CAMERA), which in addition to annotation, are able to conduct taxonomic and functional classification as well as pathway reconstruction (Wooley et al., 2010; Desai et al., 2012; Scholz et al., 2012; Ravin et al., 2015). The dependability of the taxonomic assignment and therefore the corresponding information may be decided from scores on sequence similarity and alignment coverage by quality standards or phylogenetic analyses (Peršoh, 2015).

Monitoring complex microbial communities is essential in food fermentations, in which consortia of microbial communities are naturally involved in the processes, such as fermentation and spoilage (Bokulich et al., 2016b). These technological advances, therefore represent an enormous breakthrough for microbial ecology, because metagenomics and NGS allow for in-depth insights into not only the structure, but the function of the most complex microbial communities in their natural environments (Peršoh, 2015). The following section, will therefore focus on metagenomics and how it has been applied to study the vineyard microbial communities.

VINEYARD MICROBIAL COMMUNITIES AS DERIVED FROM TARGETED SGS

SGS technologies have become the tool of choice in deciphering the vineyard and wine microbiome. Most importantly these tools have been employed in microbial surveys that sought to understand how agronomic practices influence microbial community structures and whether there are grapevine organ-specific microbial signatures. Furthermore, it is increasingly

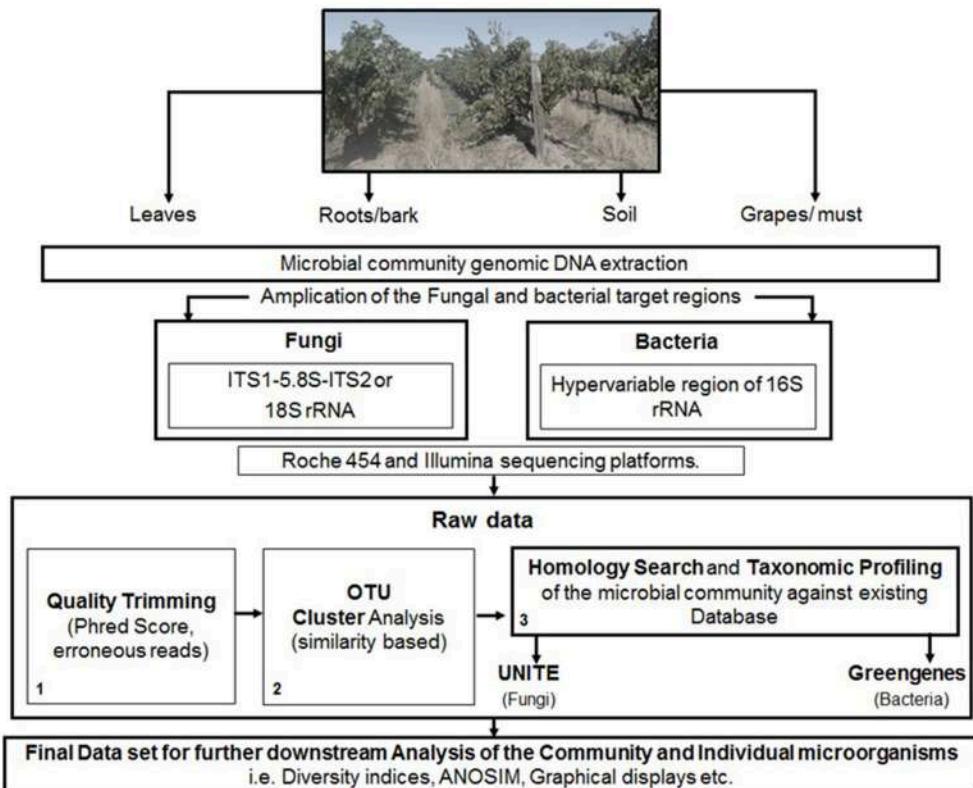


FIGURE 1 | A schematic representation of the steps involved in targeted amplicon sequencing.

becoming important to understand whether there is geographic microbial signatures that contribute to wine typicity.

BACTERIAL COMMUNITIES

Several studies have recently employed high-throughput sequencing to evaluate the bacterial communities associated with the vineyard. The most abundant phyla in vineyard soils and grapevine roots include *Proteobacteria*, *Bacteroidetes*, *Acidobacteria*, *Verrucomicrobia*, *Planctomycetes*, *Actinobacteria*, *Chloroflexi*, *Gemmatimonadetes*, and *Firmicutes* (Burns et al., 2015; Calleja-Cervantes et al., 2015; Zarraonaindia et al., 2015; Faist et al., 2016). Studies suggest that the soil microbial community composition in vineyards closely resembles that of other agricultural ecosystems and is largely structured with respect to soil properties and viticultural area (Burns et al., 2015). Furthermore, soil amendments such as fertilizer and/or compost applications can alter the relative abundances of bacterial groups (Calleja-Cervantes et al., 2015). High-throughput analysis of the grapevine phyllosphere, flowers and grape berry surface, demonstrated that the bacterial communities were predominated by *Proteobacteria* followed by *Firmicutes*, *Actinobacteria*, *Acidobacteria*, and *Bacteroidetes* (Perazzoli et al., 2014; Pinto et al., 2014, 2015; Portillo and Mas, 2016; Portillo et al., 2016). The relative abundances of the groups vary depending on the plant tissue or organ.

Dominant taxa include members of the genera *Pseudomonas*, *Sphingomonas*, *Frigoribacterium*, *Curtobacterium*, *Bacillus*, *Enterobacter*, *Acinetobacter*, *Erwinia*, *Citrobacter*, *Pantoea*, and *Methylobacterium* (Bokulich et al., 2014, 2016a; Perazzoli et al., 2014; Pinto et al., 2015; Zarraonaindia et al., 2015; Kecskeméti et al., 2016; Portillo and Mas, 2016; Portillo et al., 2016). In contrast, the endophytic community in grape berries mainly comprise *Ralstonia*, *Burkholderia*, *Pseudomonas*, *Staphylococcus*, *Mesorhizobium*, *Propionibacterium*, *Dyella*, and *Bacillus* species (Campisano et al., 2014). However, it is important to note that the bacterial community structure varies amongst grape cultivars, and is also influenced by agronomic practices (Campisano et al., 2014; Perazzoli et al., 2014; Calleja-Cervantes et al., 2015; Pinto et al., 2015; Kecskeméti et al., 2016). Furthermore, development of diseases can result in establishment of different community structures. For instance, graft unions with crown gall were shown to harbor three bacterial OTUs viz. *Agrobacterium vitis*, *Pseudomonas* sp., and *Enterobacteriaceae* sp., that were most abundant in every season, while the three most abundant OTUs in graft unions without a crown gall differed in every season suggesting that crown galls are colonized by a stable bacterial complex (Faist et al., 2016). In other studies, a higher incidence of acetic acid bacteria (AAB) was shown to develop in positive correlation with the *Botryotinia* sp. on grapevine leaves and in botrytized wine fermentations (Bokulich et al., 2012; Pinto et al., 2015). However, Portillo and Mas (2016) demonstrated

that this group of bacteria, specifically *Gluconobacter* spp., can persist at high abundance throughout wine fermentation in non-botrytized Grenache fermenting musts, only declining at the end of alcoholic fermentation. Furthermore, the population of *Gluconobacter* was shown to be highly abundant in organic pied-de-cuve Riesling fermentation compared to the conventional fermentation (Piao et al., 2015). AAB were also shown to dominate in low sulfited, uninoculated wine fermentations, compared to *Lactobacillus* and *Lactobacillaceae* that dominated SO₂-free uninoculated fermentations (Bokulich et al., 2015). Interestingly a low abundance of LAB is often reported with amplicon sequencing phylogenetic surveys (Bokulich et al., 2012; Pinto et al., 2014, 2015). Most importantly, *Oenococcus oeni* seems to be rarely encountered in grape must except in one study where it was found to be dominant in fermentations of Grenache and Carignan grapes (Portillo et al., 2016). However, several studies show that the levels of this species increase during malolactic fermentation and that in fact it is in most cases the dominant taxa (Marzano et al., 2016; Portillo et al., 2016). Other LAB often encountered include *Lactobacillus*, *Lactococcus*, *Leuconostoc*, and *Pediococcus* species (Bokulich et al., 2012, 2014; Piao et al., 2015; Pinto et al., 2015; Portillo et al., 2016).

Overall SGS have made it possible to detect bacterial species often overlooked in culture-based methods and community fingerprinting approaches such as DGGE as it can detect species that represent 0.001–1% of the total population. Furthermore, several novel genera believed to be associated with the wine habitat, including, *Candidatus Liberibacter*, *Onus*, *Wolbachia*, *Komagataeibacter*, and *Shewanella* were detected (Marzano et al., 2016; Portillo and Mas, 2016). In some cases, these rare taxa including *Methylobacterium*, *Sphingomonas*, *Acinetobacter*, *Pseudomonas*, *Wolbachia*, and *Paracoccus* could be detected until the end of alcoholic fermentation (Bokulich et al., 2012; Piao et al., 2015; Portillo and Mas, 2016). A closer look at supplementary data from various publications suggests that over 100 species are newly associated with grapevine or wine. However, since only partial sequences are used, most of the taxonomic assignments are generally reliable to genus level. Nevertheless, Table 2 shows a representation of a few species that have been identified in various studies and have been shown to persist from the vineyard environment and throughout wine fermentation. Some of the species e.g., *Methylobacterium populi* and *Sphingomonas pseudosanguinis*, were confirmed to be viable at the end of fermentation (Bokulich et al., 2012) and the populations of these genera were also shown to persist in the winery on non-fermentor surfaces (Bokulich et al., 2013b). Further research into these taxa is, however required to evaluate their possible impact in wine fermentation and/or wine quality.

FUNGAL COMMUNITIES

The fungal communities associated with grapevine have mainly been investigated in must after crushing. Overall, the fungal populations at a phylum level are very similar and mainly comprise the Ascomycota and the most abundant phylum followed by the Basidiomycota (Bokulich et al., 2014; David et al., 2014; Taylor et al., 2014; Pinto et al., 2015; Setati

et al., 2015; Kecskeméti et al., 2016). Other phyla such as the Zygomycota and Chytridiomycota are only present in low abundance. Frequently encountered genera of filamentous fungi include *Aspergillus*, *Alternaria*, *Penicillium*, *Cladosporium*, *Lewia*, *Davidiella*, *Erysiphe*, *Botrytis* and the yeast-like fungus, *Aureobasidium pullulans*, while the yeast genera include *Hanseniaspora*, *Issatchenka*, *Pichia*, *Candida*, *Rhodotorula*, *Lachancea*, *Metschnikowia*, *Cryptococcus*, *Filobasidiella*, *Sporobolomyces*, and *Torulaspora* (Bokulich et al., 2014; David et al., 2014; Taylor et al., 2014; Pinto et al., 2015; Setati et al., 2015; Wang et al., 2015; Kecskeméti et al., 2016; De Filippis et al., 2017). Generally, the SGS have revealed more filamentous fungal species than yeast species especially those associated with the grape berry surface (Tables 3, 4). These data suggest that most of the yeast genera and species are cultivable but are often missed in culture-based studies due to their presence in minor concentrations. In contrast, for the filamentous fungi, SGS reveals a diversity of possible rot associated taxa such as *Botrytis elliptica* and *Botrytis fabae*. Further studies could look into investigating the prevalence of these species and their contribution to rot.

Several studies have suggested that the microbial community associated with grapevines exhibit regional differentiation (Bokulich et al., 2014, 2016a,b; Taylor et al., 2014; Pinto et al., 2015; Wang et al., 2015). Such regional distinction has been attributed to the dominance of a few species per region. For instance, Bokulich et al. (2014) demonstrated significant association of *Aspergillus* and *Penicillium* spp. with the Chardonnay in Napa, while *Bacteroides*, *Actinobacteria*, *Saccharomycetes*, and *Erysiphe necator* were abundant in Central Coast; and *Botryotinia fuckeliana* and *Proteobacteria* in Sonoma. Similarly, Pinto et al. (2015) showed that *Lachancea* prevailed in the Alentejo appellation, while *Rhodotorula* and *Botryotinia* dominated in the Estremadura appellation, *Hanseniaspora* and *Ramularia* in Bairrada, *Lachancea* and *Rhodotorula* in Dão, *Rhodotorula* and *Erysiphe* in Douro, and *Rhodotorula* and *Alternaria* in Minho appellation. The fungal diversity associated with grapes is also influenced by agronomic practices. Most importantly, studies have shown that vineyards employing conventional, Integrated Pest management systems, Organic, Biodynamic, and Ecophyto practices harbor different fungal communities (David et al., 2014; Setati et al., 2015; Kecskeméti et al., 2016).

Overall, NGS reveal higher diversity compared to other culture-independent methods such as DGGE and qPCR (David et al., 2014; Wang et al., 2015). Furthermore, these methods have detected minor and rare species that are sometimes overlooked with culture-dependent methods and can detect non-culturable cells at the end of fermentation. For instance, some of the studies show the presence yeast genera such as *Kazachstania*, *Malassezia*, *Schizosaccharomyces*, and *Debaryomyces* which are typically at low frequency (David et al., 2014; Pinto et al., 2015; Setati et al., 2015; Grangeau et al., 2017), while cells of *Hanseniaspora* spp. have been detected at the end of fermentation (Wang et al., 2015). Similar to what has been observed with culture-dependent methods, *S. cerevisiae* is rarely encountered in grape must with NGS technologies. However, the fungal

TABLE 2 | A selection of rare bacterial species detected on grapevine leaves (L), Roots (R), Stems, and Shoots (SS), berry surface (B) and in Soil (So), Grape Marc (GM), as well as in must (M) before fermentation (BF), in the middle (MF) and at the end of the alcoholic fermentation (EF).

Genus	Species	Source	Fermentation stage	References
Acinetobacter	<i>A.baumannii</i> <i>A.calcoaceticus</i> <i>A.guillouiae</i> <i>A.johnsonii</i> <i>A.junii</i> <i>A.lwoffii</i> <i>A.rhizosphaeraeae</i>	GM, M, R, So	BF	Burns et al., 2015; Piao et al., 2015; Marzano et al., 2016; Portillo et al., 2016
Candidatus	<i>Ca. Accumulibacter unclassified</i> <i>Ca. Blochmannia floridanus</i> <i>Ca. Blochmannia pennsylvanicus</i> <i>Ca. Carsonella ruddii</i> <i>Ca. Desulfurudis audaxviator</i> <i>Ca. Liberibacter</i> <i>Ca. Pelagibacter ubique</i> <i>Ca. Phytoplasma yellows</i> <i>Ca. Sulcia muelleri</i> <i>Ca. Vesicomyosocius okutanii</i>	M	BF/MF/EF	Marzano et al., 2016; Salvetti et al., 2016
Chryseobacterium		B, GM, M, So	BF/MF/EF	Campanaro et al., 2014; Burns et al., 2015; Kecskeméti et al., 2016;
Halomonas	<i>H. desiderata</i> <i>H. elongata</i> <i>H. phoceaee</i> <i>H. rifensis</i>	B, M	BF/MF/EF	Bokulich et al., 2015; Marzano et al., 2016; Salvetti et al., 2016
Komagataeibacter	<i>K. europaeus</i> <i>K. hansenii</i> <i>K. intermedius</i> <i>K. kakiacetii</i> <i>K. maltacetii</i> <i>K. medellinensis</i> <i>K. oboediensis</i> <i>K. rhaeticus</i> <i>K. saccharivorans</i> <i>K. sucrofermentans</i> <i>K. xylinus</i>	M	BF/MF/EF	David et al., 2014; Pinto et al., 2014, 2015; Setati et al., 2015
Methylobacterium	<i>M. adhaesivum</i> <i>M. dankookense</i> <i>M. extorquens</i> <i>M. fujisawaense</i> <i>M. longum</i> <i>M. mesophilicum</i> <i>M. populi</i> <i>M. radiotolerans</i> <i>M. rhodesianum</i>	M, R, So	BF/MF/EF	Bokulich et al., 2012; Burns et al., 2015; Piao et al., 2015; Marzano et al., 2016; Portillo et al., 2016
Ralstonia	<i>R. solanacearum</i>	SS/M	BF	Campisano et al., 2014; Marzano et al., 2016; Salvetti et al., 2016

(Continued)

TABLE 2 | Continued

Genus	Species	Source	Fermentation stage	References
<i>Sphingomonas</i>	<i>S. aerolata</i> <i>S. aquatilis</i> <i>S. echinoides</i> <i>S. endophytica</i> <i>S. insulae</i> <i>S. melonis</i> <i>S. mucosissima</i> <i>S. phyllosphaerae</i> <i>S. pseudosanguinis</i> <i>S. wittichii</i> <i>S. yunnanensis</i>	B, GM, M, R, So	BF/MF	Bokulich et al., 2012; Campanaro et al., 2014; Burns et al., 2015; Piao et al., 2015; Faist et al., 2016; Kecskeméti et al., 2016; Marzano et al., 2016; Salvetti et al., 2016
<i>Wolbachia</i>	<i>W. endosymbiont</i>	M	BF/MF	Piao et al., 2015; Kecskeméti et al., 2016; Marzano et al., 2016; Portillo and Mas, 2016; Salvetti et al., 2016

community in fermenting musts tends to be less diverse toward the end of fermentation and is dominated by *Saccharomyces* spp. In some cases, where strong fermentative yeasts such as *Lachancea*, *Starmerella*, and *Schizosaccharomyces* were present at high frequency in the initial population, they persist until the end of fermentation (Pinto et al., 2015; Wang et al., 2015; Bokulich et al., 2016a). Such species have also been shown to contribute toward taxonomic discrimination between growing regions. There is also increasing evidence that there are broad taxonomic trends underlying varietal patterns. For instance, Bokulich et al. (2014) found differences in Chardonnay, Cabernet sauvignon, and Zinfandel, while Wang et al. (2015) demonstrated that Grenache and Carignan grapes harbored certain distinct taxa. Most recently, Aglianico and Greco di Tufo were also found to harbor different yeast communities (De Filippis et al., 2017). Current data show that there is conflicting outcomes regarding the relative abundances of yeast species in must depending on the methods employed. Therefore, although microbial surveys using amplicon sequencing can detect all species that are retrieved by culture-based methods, and other culture-independent methods, the quantity of certain species tends to vary. In addition, there can be variation in community composition depending on the rRNA gene target. For instance, in the study by Pinto et al. (2015) both the D2 region and the ITS-5.8S region were targeted, however, only 13.2% of the taxa were common between the two data sets. This highlights an important gap with regard to the completeness of the databases and accuracy with regard to taxonomic assignment especially at a species level. Furthermore, amplicon sequencing data still comprise significant percentages of “unclassified” or unassigned OTUs which suggests that the diversity is still to some extent under-represented. Studies evaluating fungal diversity in the vineyard remain limited. Orgiazzi et al. (2012) reported that the soil ecosystem is dominated by the genera *Penicillium* and *Cryptococcus*, the minor fungal groups are mainly dominated by *Glomeromycota* or *Chytridiomycota*. In contrast, the leaf associated microbiome is dominated by early diverging fungal lineages (*Zygomycota*) such as *Rhizopus* and *Mucor* (Pinto et al., 2014), while AMF specific fungi of the soil and grapevine are

dominated by *Glomeromycota* (Lumini et al., 2010; Bouffaud et al., 2016). However, more studies need to be performed in order to confidently elucidate the vineyard and grapevine phyllosphere microbiome.

WHOLE-METAGENOMIC SEQUENCING

Recently, Salvetti et al. (2016) employed whole genome sequencing for the first in-depth evaluation of the microbial consortium associated with Corvina berries post withering performed in two different conditions. A total of 25 bacterial phyla were detected, nine of which were common and consisted of *Acidobacteria*, *Actinobacteria*, *Cyanobacteria*, *Firmicutes*, and *Proteobacteria*; the latter was predominant, followed by *Firmicutes*, *Actinobacteria*, and *Bacteroidetes* as reported by Pinto et al. (2014) and Zarraonaindia et al. (2015), who both employed target metagenomics strategies. The class *Gammaproteobacteria* was dominant, which was further represented by *Pseudomonadaceae* in high abundances in the traditional withering and *Enterobacteriaceae* in accelerated withering. Furthermore, both genera *Carnobacterium* and *Enterococcus* previously identified as grape associated by Pinto et al. (2015) was detected using the whole genome sequencing approach. Also, evaluating the eukaryotic community, they reported that *Ascomycota* was the dominant phylum, more specifically the class *Eurotiomycetes*, specifically genera belonging to *Aspergillus* and *Penicillium*, followed by *Sordariomycetes* and *Dothideomycetes*. However, common yeast such as *Aureobasidium*, *Cryptococcus*, *Hanseniaspora*, *Metschnikowia*, and *Sporobolomyces* which are regularly detected in targeted strategies were not detected.

Beyond providing the inventory of the vineyard, whole metagenomic analysis provides the functional information for the evaluated microbiome. For instance, information regarding defense, amino acid metabolism, transport, transcription and carbohydrate metabolism, potentially allowing a greater comparison to be drawn than the assumed microbial diversity and composition (Campanaro et al., 2014; Salvetti et al., 2016).

TABLE 3 | Filamentous fungi detected on grapevine leaves (L), berry surface (B) and in must (M) before fermentation (BF), in the middle (MF) and at the end of the alcoholic fermentation (EF).

Genus	Species	Source	Stage	References
FILAMENTOUS FUNGI				
<i>Albugo</i>	<i>A. laibachii</i>	B		Kecskeméti et al., 2016
<i>Ascochyta</i>	<i>A. fabae, A. rabiei</i>	M		Setati et al., 2015
<i>Botrytis</i>	<i>Bot. elliptica Bot. fabae</i>	B/M		Kecskeméti et al., 2016; Setati et al., 2015
<i>Cadophora</i>	<i>C. luteo-olivacea</i>	B		Kecskeméti et al., 2016
<i>Catelunostroma</i>	<i>C. protearum</i>	B		Kecskeméti et al., 2016
<i>Chloroscypha</i>	<i>C. enterochroma</i>	B		Kecskeméti et al., 2016
<i>Cladosporium</i>	<i>C. cucumerinum</i>	B/M	BF/MF/EF	Bokulich et al., 2014, 2016a; Taylor et al., 2014; De Filippis et al., 2017; Graneteau et al., 2017; Kecskeméti et al., 2016; Setati et al., 2015
	<i>C. exasperatum</i>			
	<i>C. flabelliforme</i>			
	<i>C. perangustum</i>			
<i>Cytospora</i>	<i>C. sacculus</i>	M	BF	Wang et al., 2015
<i>Didymella</i>	<i>D. exitialis D. fabae</i>	B		Kecskeméti et al., 2016
<i>Gigaspora</i>	<i>G. margarita</i>	B		Kecskeméti et al., 2016
<i>Glonium</i>	<i>G. pusillum</i>	B		Kecskeméti et al., 2016
<i>Haplographium</i>	<i>H. catenatum</i>	B		Kecskeméti et al., 2016
<i>Holtermannia</i>	<i>H. corniformis</i>	B		Kecskeméti et al., 2016
<i>Hypholoma</i>	<i>H. fasciculare</i>	B		Kecskeméti et al., 2016
<i>Kabatiella</i>	<i>K. microsticta</i>	M		Setati et al., 2015
<i>Mycosphaerella</i>	<i>M. milleri</i>	M		De Filippis et al., 2017
<i>Pandora</i>	<i>P. neoaphidis</i>	L		Pinto et al., 2014
<i>Peniosphora</i>	<i>P. aurantiaca</i>	B		Kecskeméti et al., 2016
	<i>P. incarnate</i>			
<i>Piptoporus</i>	<i>P. betulinus</i>	B		Kecskeméti et al., 2016
<i>Puccinia</i>	<i>P. punctiformis</i>	L/B		Pinto et al., 2014; Kecskeméti et al., 2016
<i>Sarocladium</i>	<i>S. strictum</i>			Kecskeméti et al., 2016
<i>Sclerotinia</i>	<i>S. subarctica</i>	B/M	BF	David et al., 2014; Kecskeméti et al., 2016; Salvetti et al., 2016
<i>Sebacina</i>	<i>S. vermicifera</i>	B		Kecskeméti et al., 2016
<i>Sphaeropsis</i>	<i>S. sapinea</i>	B		Kecskeméti et al., 2016
<i>Stephanonectaria</i>	<i>S. keithii</i>	B		Kecskeméti et al., 2016
<i>Sydowia</i>	<i>S. polyspora</i>	B		Kecskeméti et al., 2016
<i>Veluticeps</i>	<i>V. berkeleyi</i>	B		Kecskeméti et al., 2016
<i>Vuilleminia</i>	<i>V. comedens</i>	B		Kecskeméti et al., 2016
<i>Zoophthora</i>	<i>Z. radicans</i>	L		Pinto et al., 2014

TABLE 4 | Yeasts detected on grapevine leaves (L), berry surface (B) and in must (M) before fermentation (BF), in the middle (MF) and at the end of the alcoholic fermentation (EF).

Genus	Species	Source	Fermentation stage	References
<i>Cryptococcus</i>	<i>C. tephrensis</i>	L/B/M	BF/MF	Bokulich et al., 2014; David et al., 2014; Taylor et al., 2014; Graneteau et al., 2017; Kecskeméti et al., 2016; Setati et al., 2015; De Filippis et al., 2017
	<i>C. chenovii</i>			
	<i>C. stepposus</i>			
<i>Filobasidium</i>	<i>F. floriforme</i>	B		Kecskeméti et al., 2016
<i>Hanseniaspora</i>	<i>H. thailandica</i>	L/M	BF/MF/EF	Wang et al., 2015
<i>Rhodotorula</i>	<i>R. fujisanensis</i>	L/M	BF	David et al., 2014; Pinto et al., 2014, 2015; Setati et al., 2015
<i>Schizosaccharomyces</i>	<i>S. japonicus</i>	M	BF/MF/EF	Pinto et al., 2015
<i>Sclerostagonospora</i>	<i>Scl. opuntiae</i>	M	BF	Bokulich et al., 2014
<i>Sporobolomyces</i>	<i>S. coprosmae</i>	M	BF/MF	David et al., 2014; Setati et al., 2015; De Filippis et al., 2017
	<i>S. oryzicola</i>			

CONCLUSION

The invaluable contribution of metagenomic approaches in deciphering the vineyard microbiome and its application provides great insights in the microbial community composition and structure of both bacteria and fungi. Metagenomic approaches provide an opportunity to study the entire microbial population and not just one group as typically done with culture-based methods. Consequently, it has been possible to assess the population dynamics during fermentation, to evaluate grapevine disease complexes and unravel unique microbial signatures present in grapevine and not in neighboring plants. Furthermore, these approaches have been valuable in understanding the influence of vineyard management practices on the grapevine microbiome. Based on the existing research papers, it appears as though the grapevine microbiome is less complex compared to other ecosystems such as soil and that a large proportion of the yeast species associated with the grape and wine environment are cultivable. This is advantageous as the species can then be evaluated for potential genes, enzymes etc. that can be of importance for winemaking. However, most of the studies show that a significant percentage of the sequence data (OTU's) remained unassigned. This problem highlights

existing challenges with sequence databases used for taxonomic assignment that are not complete and for this technology to be furthered in future means that the expansion of the databases are crucial. Nevertheless, based on existing data, sequence-based methods reveal similar fungal species compared to culture-dependent methods, especially regarding the yeasts which are relevant in wine fermentation. The discovery of new species associated with the grape and wine microbiome holds tremendous potential to mine them for novel properties that would improve wine fermentation, aroma and style.

AUTHOR CONTRIBUTIONS

HM wrote the first draft of the review; MdT and MS proofed the drafts and finalized the review.

FUNDING

This work was funded by the National Research Foundation (NRF) [grant number 101998] and Winetech SU IWBT 16/02. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the funding agencies.

REFERENCES

- Abbasian, F., Lockington, R., Mallavarapu, M., and Naidu, R. (2015a). A pyrosequencing-based analysis of microbial diversity governed by ecological conditions in the Winogradsky column. *World J. Microbiol. Biotechnol.* 31, 1115–1126. doi: 10.1007/s11274-015-1861-y
- Abbasian, F., Lockington, R., Megharaj, M., and Naidu, R. (2015b). The integration of sequencing and bioinformatics in metagenomics. *Rev. Environ. Sci. Biotechnol.* 14, 357–383. doi: 10.1007/s11157-015-9365-7
- Alessandria, V., Giacosa, S., Campolongo, S., Rolle, L., Rantsiou, K., and Cocolin, L. (2013). Yeast population diversity on grapes during on-vine withering and their dynamics in natural and inoculated fermentations in the production of icewines. *Food Res. Int.* 54, 139–147. doi: 10.1016/j.foodres.2013.06.018
- Andorrà, I., Landi, S., Mas, A., Esteve-Zarzoso, B., and Guillamón, J. M. (2010). Effect of fermentation temperature on microbial population evolution using culture-independent and dependent techniques. *Food Res. Int.* 43, 773–779. doi: 10.1016/j.foodres.2009.11.014
- Andorrà, I., Landi, S., Mas, A., Guillamón, J. M., and Esteve-Zarzoso, B. (2008). Effect of oenological practices on microbial populations using culture-independent techniques. *Food Microbiol.* 25, 849–856. doi: 10.1016/j.fm.2008.05.005
- Arteau, M., Labrie, S., and Roy, D. (2010). Terminal-restriction fragment length polymorphism and automated ribosomal intergenic spacer analysis profiling of fungal communities in Camembert cheese. *Int. Dairy J.* 20, 545–554. doi: 10.1016/j.idairyj.2010.02.006
- Balázs, M., Rónavári, A., Németh, A., Bihari, Z., Rutkai, E., Bartos, P., et al. (2013). Effect of DNA polymerases on PCR-DGGE patterns. *Int. Biodeterior. Biodegrad.* 84, 244–249. doi: 10.1016/j.ibiod.2012.05.011
- Barata, A., Malfeito-Ferreira, M., and Loureiro, V. (2012). The microbial ecology of wine grape berries. *Int. J. Food Microbiol.* 153, 243–259. doi: 10.1016/j.ijfoodmicro.2011.11.025
- Barnett, J. A. (2003). Beginnings of microbiology and biochemistry: the contribution of yeast research. *Microbiology* 149, 557–567. doi: 10.1099/mic.0.26089-0
- Bleidorn, C. (2015). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Syst. Biodivers.* 2000, 1–8. doi: 10.1080/14772000.2015.1099575
- Bokulich, N. A., Collins, T. S., Masarweh, C., Allen, G., Heymann, H., Ebeler, S. E., et al. (2016a). Associations among wine grape microbiome, metabolome, and fermentation behavior suggest microbial contribution to regional wine characteristics. *MBio* 7, 1–12. doi: 10.1128/mBio.00631-16
- Bokulich, N. A., Joseph, C. M. L., Allen, G., Benson, A. K., and Mills, D. A. (2012). Next-generation sequencing reveals significant bacterial diversity of botrytized wine. *PLoS ONE* 7:e36357. doi: 10.1371/journal.pone.0036357
- Bokulich, N. A., Lewis, Z. T., Boundy-Mills, K., and Mills, D. A. (2016b). A new perspective on microbial landscapes within food production. *Curr. Opin. Biotechnol.* 37, 182–189. doi: 10.1016/j.copbio.2015.12.008
- Bokulich, N. A., and Mills, D. A. (2013). Improved selection of internal transcribed spacer-specific primers enables quantitative, ultra-high-throughput profiling of fungal communities. *Appl. Environ. Microbiol.* 79, 2519–2526.
- Bokulich, N. A., Ohta, M., Richardson, P. M., and Mills, D. A. (2013b). Monitoring seasonal changes in winery-resident microbiota. *PLoS ONE* 8:e66437. doi: 10.1371/journal.pone.0066437
- Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, I., Knight, R., et al. (2013a). Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi: 10.1038/nmeth.2276
- Bokulich, N. A., Swadener, M., Sakamoto, K., Mills, D. A., and Bisson, L. F. (2015). Sulfur dioxide treatment alters wine microbial diversity and fermentation. Progression in a dose-dependent fashion. *Am. J. Enol. Vitic.* 66, 73–79. doi: 10.5344/ajev.2014.14096
- Bokulich, N. A., Thorngate, J. H., Richardson, P. M., and Mills, D. A. (2014). Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. *Proc. Natl. Acad. Sci. U.S.A.* 111, E139–E148. doi: 10.1073/pnas.1317377111
- Bouffaud, M. L., Bernaud, E., Colombet, A., Van Tuinen, D., Wipf, D., and Redecker, D. (2016). Regional-scale analysis of arbuscular mycorrhizal fungi: the case of Burgundy vineyards. *J. Int. Sci. Vigne Vin.* 50, 1–8. doi: 10.20870/oeno-one.2016.50.1.49
- Brežná, B., Zenišová, K., Chovanová, K., Chebeňová, V., Kraková, L., Kuchta, T., et al. (2010). Evaluation of fungal and yeast diversity in Slovakian wine-related microbial communities. *Antonie Van Leeuwenhoek* 98, 519–529. doi: 10.1007/s10482-010-9469-6

- Burns, K. N., Kluepfel, D. A., Strauss, S. L., Bokulich, N. A., Cantu, D., and Steenwerth, K. L. (2015). Vineyard soil bacterial diversity and composition revealed by 16S rRNA genes: differentiation by geographic features. *Soil Biol. Biochem.* 91, 232–247. doi: 10.1016/j.soilbio.2015.09.002
- Calleja-Cervantes, M. E., Menéndez, S., Fernández-González, A. J., Irigoyen, I., Cibrián-Sabalza, J. F., Toro, N., et al. (2015). Changes in soil nutrient content and bacterial community after 12 years of organic amendment application to a vineyard. *Eur. J. Soil Sci.* 66, 802–812. doi: 10.1111/ejss.12261
- Cameron, M., Siebrito, L., du Toit, M., and Witthuhn, R. C. (2013). PCR-based DGGE fingerprinting and identification of the microbial population in South African red grape must and wine. *J. Int. Sci. Vigne Vin.* 47, 47–54. doi: 10.20870/oeno-one.2013.47.1.1531
- Campanaro, S., Treu, L., Vendramin, V., Bovo, B., Giacomini, A., and Corich, V. (2014). Metagenomic analysis of the microbial community in fermented grape marc reveals that *Lactobacillus fabifermentans* is one of the dominant species: insights into its genome structure. *Appl. Microbiol. Biotechnol.* 98:60156037. doi: 10.1007/s00253-014-5795-3
- Campisano, A., Antonielli, L., Pancher, M., Yousaf, S., Pindo, M., and Pertot, I. (2014). Bacterial endophytic communities in the grapevine depend on pest management. *PLoS ONE* 9:e112763. doi: 10.1371/journal.pone.0112763
- Capozzi, V., Di Toro, M. R., Grieco, F., Michelotti, V., Salma, M., Lamontanara, A., et al. (2016). Viable but not Culturable (VBNC) state of *Brettanomyces bruxellensis* in wine: new insights on the molecular basis of VBNC behavior using a transcriptomic approach. *Food Microbiol.* 59, 196–204. doi: 10.1016/j.fm.2016.06.007
- Chovanová, K., Kraková, L., Ženíšková, K., Turcovská, V., Brežná, B., Kuchta, T., et al. (2011). Selection and identification of autochthonous yeasts in Slovakian wine samples using a rapid and reliable three-step approach. *Lett. Appl. Microbiol.* 53, 231–237. doi: 10.1111/j.1472-765X.2011.03097.x
- Cocolin, L., Heisey, A., and Mills, D. A. (2001). Direct identification of the indigenous yeasts in commercial wine fermentations. *Am. J. Enol. Vitic.* 52, 49–53.
- Cocolin, L., Alessandria, V., Dolci, P., Gorra, R., and Rantsiou, K. (2013). Culture independent methods to assess the diversity and dynamics of microbiota during food fermentation. *Int. J. Food Microbiol.* 167, 29–43. doi: 10.1016/j.ijfoodmicro.2013.05.008
- Comitini, F., Capece, A., Ciani, M., and Romano, P. (2017). New insights on the use of wine yeasts. *Curr. Opin. Food Sci.* 13, 44–49. doi: 10.1016/j.cofs.2017.02.005
- David, V., Terrat, S., Herzine, K., Claisse, O., Rousseaux, S., Tourdot-Maréchal, R., et al. (2014). High-throughput sequencing of amplicons for monitoring yeast biodiversity in must and during alcoholic fermentation. *J. Ind. Microbiol. Biotechnol.* 41, 811–821. doi: 10.1007/s10295-014-1427-2
- De Filippis, F., La Storia, A., and Blaiotta, G. (2017). Monitoring the mycobiota during Greco di Tufo and Anglianico wine fermentation by 18S rRNA gene sequencing. *Food Microbiol.* 63, 117–122. doi: 10.1016/j.fm.2016.11.010
- Desai, N., Antonopoulos, D., Gilbert, J. A., Glass, E. M., and Meyer, F. (2012). From genomics to metagenomics. *Curr. Opin. Biotechnol.* 23, 72–76. doi: 10.1016/j.copbio.2011.12.017
- Díaz-Sánchez, S., Hanning, I., Pendleton, S., and D'Souza, D. (2013). Next-generation sequencing: the future of molecular genetics in poultry production and food safety. *Poult. Sci.* 92, 562–572. doi: 10.3382/ps.2012-02741
- Di Maro, E., Ercolini, D., and Coppola, S. (2007). Yeast dynamics during spontaneous wine fermentation of the Catalanesca grape. *Int. J. Food Microbiol.* 117, 201–210. doi: 10.1016/j.ijfoodmicro.2007.04.007
- Divilo, B., and Lonvaud-Funel, A. (2005). Evidence for viable but noncultivable yeasts in *Botrytis*-affected wine. *J. Appl. Microbiol.* 99, 85–93. doi: 10.1111/j.1365-2672.2005.02578.x
- Esteve-Zarzoso, B. (1999). Identification of yeasts by RFLP analysis of the 5.8 S rRNA gene and the two ribosomal internal transcribed spacers. *Int. J. Syst. Bacteriol.* 49, 329–337. doi: 10.1099/00207713-49-1-329
- Faist, H., Keller, A., Hentschel, U., and Deeken, R. (2016). Grapevine (*Vitis vinifera*) crown galls host distinct microbiota. *Appl. Environ. Microbiol.* 82, 5542–5552. doi: 10.1128/AEM.01131-16
- Fasoli, S., Marzotto, M., Rizzotti, L., Rossi, F., Dellaglio, F., and Torriani, S. (2003). Bacterial composition of commercial probiotic products as evaluated by PCR-DGGE analysis. *Int. J. Food Microbiol.* 82, 59–70. doi: 10.1016/S0168-1605(02)00259-3
- Fernández-Espinar, M. T., López, V., Ramón, D., Bartra, E., and Querol, A. (2001). Study of the authenticity of commercial wine yeast strains by molecular techniques. *Int. J. Food Microbiol.* 70, 1–10. doi: 10.1016/S0168-1605(01)00502-5
- Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., et al. (2015). Sequencing and beyond: integrating molecular “omics” for microbial community profiling. *Nat. Rev. Microbiol.* 13, 360–372. doi: 10.1038/nrmicro3451
- Ghosh, S., Bagheri, B., Morgan, H. H., Divol, B., and Setati, M. E. (2015). Assessment of wine microbial diversity using ARISA and cultivation-based methods. *Ann. Microbiol.* 65, 1833–1840. doi: 10.1007/s13213-014-1021-x
- Gilbert, J. A., van der Lelie, D., and Zarraonaindia, I. (2014). Microbial terroir for wine grapes. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5–6. doi: 10.1073/pnas.1320471110
- Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49
- Grangeau, C., Roullier-Gall, C., Rousseaux, S., Gugeon, R. D., Schmitt-Kopplin, P., Alexandre, H., et al. (2017). Wine microbiology is driven by vineyard and winery anthropogenic factors. *Microb. Biotechnol.* 10, 354–370. doi: 10.1111/1751-7915.12428
- Grube, M., Schmid, F., and Berg, G. (2011). Black fungi and associated bacterial communities in the phyllosphere of grapevine. *Fungal Biol.* 115, 978–986. doi: 10.1016/j.funbio.2011.04.004
- Guzzo, R., Nicolini, G., Nardin, T., Malacarne, M., and Larcher, R. (2014). Survey about the microbiological features, the oenological performance and the influence on the character of wine of active dry yeast employed as starters of wine fermentation. *Int. J. Food Sci. Technol.* 49, 2142–2148. doi: 10.1111/ijfs.12610
- Heather, J. M., and Chain, B. (2016). The sequence of sequencers: the history of sequencing DNA. *Genomics* 107, 1–8. doi: 10.1016/j.ygeno.2015.11.003
- Holland, T. C., Bowen, P. A., Bogdanoff, C. P., Lowery, T. D., Shaposhnikova, O., Smith, S., et al. (2016). Evaluating the diversity of soil microbial communities in vineyards relative to adjacent native ecosystems. *Agric. Ecosyst. Environ. Appl. Soil Ecol.* 100, 91–103. doi: 10.1016/j.apsoil.2015.12.001
- Holland, T. C., Bowen, P., Bogdanoff, C., and Hart, M. (2014). Arbuscular mycorrhizal fungal communities associated with *Vitis vinifera* vines under different frequencies of irrigation. *Am. J. Enol. Vitic.* 65, 222–229. doi: 10.5344/ajev.2014.13101
- Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8:R143. doi: 10.1186/gb-2007-8-7-r143
- Jolly, N. P., Varela, C., and Pretorius, I. S. (2014). Not your ordinary yeast: non-Saccharomyces yeasts in wine production uncovered. *FEMS Yeast Res.* 14, 215–237. doi: 10.1111/1567-1364.12111
- Justé, A., Thomma, B. P. H. J., and Lievens, B. (2008). Recent advances in molecular techniques to study microbial communities in food-associated matrices and processes. *Food Microbiol.* 25, 745–761. doi: 10.1016/j.fm.2008.04.009
- Kecskeméti, E., Berkemann-Löhnertz, B., and Reineke, A. (2016). Are epiphytic microbial communities in the carposphere of ripening grape clusters (*Vitis vinifera* L.) different between conventional, organic, and biodynamic grapes? *PLoS ONE* 11:e0160852. doi: 10.1371/journal.pone.0160852
- Köljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., et al. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* 22, 5271–5277. doi: 10.1111/mec.12481
- Kovacs, A., Yacob, K., and Gophna, U. (2010). A systematic assessment of automated ribosomal intergenic spacer analysis (ARISA) as a tool for estimating bacterial richness. *Res. Microbiol.* 161, 192–197. doi: 10.1016/j.resmic.2010.01.006
- Kraková, L., Chovanová, K., Ženíšková, K., Turcovská, V., Brežná, B., Kuchta, T., et al. (2012). Yeast diversity investigation of wine-related samples from two different Slovakian wine-producing areas through a multistep procedure. *LWT Food Sci Technol.* 46, 406–411. doi: 10.1016/j.lwt.2011.12.010
- Kriseman, J., Busick, C., Szelingher, S., and Dinu, V. (2010). BING: biomedical informatics pipeline for next generation sequencing. *J. Biomed. Inform.* 43, 428–434. doi: 10.1016/j.jbi.2009.11.003
- Leveau, J. H. J., and Tech, J. J. (2011). Grapevine microbiomics: bacterial diversity on grape leaves and berries revealed by high-throughput

- sequence analysis of 16S rRNA amplicons. *Acta Hortic.* 905, 31–42. doi: 10.17660/ActaHortic.2011.905.2
- Lu, Y., Huang, D., Lee, P. R., and Liu, S. Q. (2016). Assessment of volatile and non-volatile compounds in durian wines fermented with four commercial non-Saccharomyces yeasts. *J. Sci. Food Agric.* 96, 1511–1521. doi: 10.1002/jsfa.7253
- Lumini, E., Orgiazzi, A., Borriello, R., Bonfante, P., and Bianciotto, V. (2010). Disclosing arbuscular mycorrhizal fungal biodiversity in soil through a land-use gradient using a pyrosequencing approach. *Environ. Microbiol.* 12, 2165–2179. doi: 10.1111/j.1462-2920.2009.02099.x
- Lv, X. C., Huang, R. L., Chen, F., Zhang, W., Rao, P. F., and Ni, L. (2013). Bacterial community dynamics during the traditional brewing of Wuyi Hong Qu glutinous rice wine as determined by culture-independent methods. *Food Control.* 34, 300–306. doi: 10.1016/j.foodcont.2013.05.003
- Ma, Q., Qu, Y., Shen, W., Zhang, Z., Wang, J., Liu, Z., et al. (2015). Bacterial community compositions of coking wastewater treatment plants in steel industry revealed by Illumina high-throughput sequencing. *Bioresour. Technol.* 179, 436–443. doi: 10.1016/j.biortech.2014.12.041
- Martins, G., Miot-Sertier, C., Lauga, B., Claisse, O., Lonvaud-Funel, A., Soulas, G., et al. (2012). Grape berry bacterial microbiota: impact of the ripening process and the farming system. *Int. J. Food Microbiol.* 158, 93–100. doi: 10.1016/j.ijfoodmicro.2012.06.013
- Martins, G., Vallance, J., Mercier, A., Albertin, W., Stamatopoulos, P., Rey, P., et al. (2014). Influence of the farming system on the epiphytic yeasts and yeast-like fungi colonizing grape berries during the ripening process. *Int. J. Food Microbiol.* 177, 21–28. doi: 10.1016/j.ijfoodmicro.2014.02.002
- Marzano, M., Fosso, B., Manzari, C., Grieco, F., Intrantuovo, M., Cozzi, G., et al. (2016). Complexity and dynamics of the winemaking bacterial communities in berries, musts, and wines from apulian grape cultivars through time and space. *PLoS ONE* 11:e0157383. doi: 10.1371/journal.pone.0157383
- Morales, P., Rojas, V., Quirós, M., and Gonzalez, R. (2015). The impact of oxygen on the final alcohol content of wine fermented by a mixed starter culture. *Appl. Microbiol. Biotechnol.* 99, 3993–4003.
- Morozova, O., and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92, 255–264. doi: 10.1016/j.ygeno.2008.07.001
- Nocker, A., Burr, M., and Camper, A. K. (2007). Genotypic microbial community profiling: a critical technical review. *Microb. Ecol.* 54, 276–289. doi: 10.1007/s00248-006-9199-5
- Orgiazzi, A., Lumini, E., Nilsson, R. H., Girlanda, M., Vizzini, A., Bonfante, P., et al. (2012). Unravelling soil fungal communities from different Mediterranean land-use backgrounds. *PLoS ONE* 7:e34847. doi: 10.1371/journal.pone.0034847
- Padilla, B., Gil, J. V., and Manzanares, P. (2016). Past and future of non-saccharomyces yeast: from spoilage microorganisms to biotechnological tools for improving wine aroma complexity. *Front. Microbiol.* 7:411. doi: 10.3389/fmicb.2016.00411
- Panchar, M., Ceol, M., Corneo, P. E., Longa, C. M. O., Yousaf, S., Pertot, I., et al. (2012). Fungal endophytic communities in grapevines (*Vitis vinifera* L.) respond to crop management. *Appl. Environ. Microbiol.* 78, 4308–4317. doi: 10.1128/AEM.07655-11
- Perazzoli, M., Antonielli, L., Storari, M., Puopolo, G., Panchar, M., Giovannini, O., et al. (2014). Resilience of the natural phyllosphere microbiota of the grapevine to chemical and biological pesticides. *Appl. Environ. Microbiol.* 80, 3585–3596. doi: 10.1128/AEM.00415-14
- Peršoh, D. (2015). Plant-associated fungal communities in the light of meta'omics. *Fungal Divers.* 75, 1–25. doi: 10.1007/s13225-015-0334-9
- Pettersson, E., Lundeberg, J., and Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics* 93, 105–111. doi: 10.1016/j.ygeno.2008.10.003
- Piao, H., Hawley, E., Kopf, S., DeScenzo, R., Sealock, S., Henick-Kling, T., et al. (2015). Insights into the bacterial community and its temporal succession during the fermentation of wine grapes. *Front. Microbiol.* 6:809. doi: 10.3389/fmicb.2015.00809
- Pinto, C., Pinho, D., Cardoso, R., Custódio, V., Fernandes, J., Sousa, S., et al. (2015). Wine fermentation microbiome: a landscape from different Portuguese wine appellations. *Front. Microbiol.* 6:905. doi: 10.3389/fmicb.2015.00905
- Pinto, C., Pinho, D., Sousa, S., Pinheiro, M., Egas, C., and Gomes, A. C. (2014). Unravelling the diversity of grapevine microbiome. *PLoS ONE* 9:e85622. doi: 10.1371/journal.pone.0085622
- Portillo, M. C., Franquès, J., Araque, I., Reguant, C., and Bordons, A. (2016). Bacterial diversity of Grenache and Carignan grape surface from different vineyards at Priorat wine region (Catalonia, Spain). *Int. J. Food Microbiol.* 219, 56–63. doi: 10.1016/j.ijfoodmicro.2015.12.002
- Portillo, M. C., and Mas, A. (2016). Analysis of microbial diversity and dynamics during wine fermentation of Grenache grape variety by high-throughput barcoding. *LWT Food Sci. Technol.* 72, 317–321. doi: 10.1016/j.lwt.2016.05.009
- Prakitchaiwattana, C. J., Fleet, G. H., and Heard, G. M. (2004). Application and evaluation of denaturing gradient gel electrophoresis to analyse the yeast ecology of wine grapes. *FEMS Yeast Res.* 4, 865–877. doi: 10.1016/j.femsyr.2004.05.004
- Ravin, N. V., Mardanov, A. V., and Skryabin, K. G. (2015). Metagenomics as a tool for the investigation of uncultured microorganisms. *Russ. J. Genet.* 51, 431–439. doi: 10.1134/S1022795415050063
- Renouf, V., Claisse, O., and Lonvaud-Funel, A. (2005). Understanding the microbial ecosystem on the grape berry surface through numeration and identification of yeast and bacteria. *Aust. J. Grape Wine Res.* 11, 316–327. doi: 10.1111/j.1755-0238.2005.tb00031.x
- Renouf, V., Claisse, O., and Lonvaud-Funel, A. (2006a). *rpoB* gene: a target for identification of LAB cocci by PCR-DGGE and melting curves analyses in real time PCR. *J. Microbiol. Methods* 67, 162–170. doi: 10.1016/j.mimet.2006.03.008
- Renouf, V., Claisse, O., and Lonvaud-Funel, A. (2007). Inventory and monitoring of wine microbial consortia. *Appl. Microbiol. Biotechnol.* 75, 149–164. doi: 10.1007/s00253-006-0798-3
- Renouf, V., Claisse, O., Miot-Sertier, C., and Lonvaud-Funel, A. (2006b). Lactic acid bacteria evolution during winemaking: use of *rpoB* gene as a target for PCR-DGGE analysis. *Food Microbiol.* 23, 136–145. doi: 10.1016/j.fm.2005.01.019
- Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Mol. Cell* 58, 586–597. doi: 10.1016/j.molcel.2015.05.004
- Rousseaux, S., Diguta, C. F., Radoi-Matei, F., Alexandre, H., and Guilloux-Bénatier, M. (2014). Non-Botrytis grape-rotting fungi responsible for earthy and moldy off-flavors and mycotoxins. *Food Microbiol.* 38, 104–121. doi: 10.1016/j.fm.2013.08.013
- Salma, M., Rousseaux, S., Sequeira-Le Grand, A., Divol, B., and Alexandre, H. (2013). Characterization of the Viable but Nonculturable (VBNC) state in *Saccharomyces cerevisiae*. *PLoS ONE* 8:e77600. doi: 10.1371/journal.pone.0077600
- Salvetti, E., Campanaro, S., Campedelli, I., Fracchetti, F., Gobbi, A., Tornielli, G. B., et al. (2016). Whole-metagenome-sequencing-based community profiles of *Vitis vinifera* L. cv. Corvina berries withered in two post-harvest conditions. *Front. Microbiol.* 7:937. doi: 10.3389/fmicb.2016.00937
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463–5467. doi: 10.1073/pnas.74.12.5463
- Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Hum. Mol. Genet.* 19, 227–240. doi: 10.1093/hmg/ddq416
- Schmid, F., Moser, G., Müller, H., and Berg, G. (2011). Functional and structural microbial diversity in organic and conventional viticulture: organic farming benefits natural biocontrol agents. *Appl. Environ. Microbiol.* 77, 2188–2191. doi: 10.1128/AEM.02187-10
- Scholz, M. B., Lo, C. C., and Chain, P. S. G. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotechnol.* 23, 9–15. doi: 10.1016/j.copbio.2011.11.013
- Setati, M. E., Jacobson, D., Andong, U. C., and Bauer, F. (2012). The vineyard yeast microbiome, a mixed model microbial map. *PLoS ONE* 7:e52609. doi: 10.1371/journal.pone.0052609
- Setati, M. E., Jacobson, D., and Bauer, F. F. (2015). Sequence-based analysis of the *Vitis vinifera* L. cv Cabernet Sauvignon grape must mycobiome in three South African vineyards employing distinct agronomic systems. *Front. Microbiol.* 6:1358. doi: 10.3389/fmicb.2015.01358
- Shi, Y., Lou, K., Li, C., Wang, L., Zhao, Z., Zhao, S., et al. (2015). Illumina-based analysis of bacterial diversity related to halophytes *Salicornia europaea* and *Suaeda aralocaspica*. *J. Microbiol.* 53, 678–685. doi: 10.1007/s12275-015-5080-x
- Slabbert, E., Kongor, R. Y., Esler, K. J., and Jacobs, K. (2010). Microbial diversity and community structure in Fynbos soil. *Mol. Ecol.* 19, 1031–1041. doi: 10.1111/j.1365-294X.2009.04517.x

- Solieri, L., and Giudici, P. (2008). Yeasts associated to traditional balsamic vinegar: ecological and technological features. *Int. J. Food Microbiol.* 125, 36–45. doi: 10.1016/j.ijfoodmicro.2007.06.022
- Stefanini, I., Albanese, D., Cavazza, A., Franciosi, E., De Filippo, C., Donati, C., et al. (2016). Dynamic changes in microbiota and mycobiota during spontaneous “Vino Santo Trentino” fermentation. *Microb. Biotechnol.* 9, 195–208. doi: 10.1111/1751-7915.12337
- Sun, Y., and Liu, Y. (2014). Investigating of yeast species in wine fermentation using terminal restriction fragment length polymorphism method. *Food Microbiol.* 38, 201–207. doi: 10.1016/j.fm.2013.09.001
- Taylor, M. W., Tsai, P., Anfang, N., Ross, H. A., and Goddard, M. R. (2014). Pyrosequencing reveals regional differences in fruit-associated fungal communities. *Environ. Microbiol.* 16, 2848–2858. doi: 10.1111/1462-2920.12456
- Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* 2:3. doi: 10.1186/2042-5783-2-3
- Türktaş, M., Kurtoglu, K. Y., Dorado, G., Zhang, B., Hernandez, P., and Ünver, T. (2015). Sequencing of plant genomes - a review. *Turk. J. Agric. Forest.* 39, 361–376. doi: 10.3906/tar-1409-93
- Wang, C., Garcia-Fernández, D., Mas, A., and Esteve-Zarzoso, B. (2015). Fungal diversity in grape must and wine fermentation assessed by massive sequencing, quantitative PCR and DGGE. *Front. Microbiol.* 6:1156. doi: 10.3389/fmicb.2015.01156
- Wang, Z. K., Yang, Y. S., Stefka, A. T., Sun, G., and Peng, L. H. (2014). Review article: fungal microbiota and digestive diseases. *Aliment. Pharmacol. Ther.* 39, 751–766. doi: 10.1111/apt.12665
- Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput. Biol.* 6:e1000667. doi: 10.1371/journal.pcbi.1000667
- Zarraonaindia, I., Owens, S. M. S., Weisenhorn, P., West, K., Hampton-Marcell, J., Lax, S., et al. (2015). The soil microbiome influences grapevine-associated microbiota. *Mbio* 6, e02527–e02514. doi: 10.1128/mBio.02527-14
- Ženišová, K., Chovanová, K., Chebeňová-Turcovská, V., Godálová, Z., Kraková, L., Kuchta, T., et al. (2014). Mapping of wine yeast and fungal diversity in the Small Carpathian wine-producing region (Slovakia): evaluation of phenotypic, genotypic and culture-independent approaches. *Ann. Microbiol.* 64, 1819–1828. doi: 10.1007/s13213-014-0827-x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Morgan, du Toit and Setati. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



From Vineyard Soil to Wine Fermentation: Microbiome Approximations to Explain the “terroir” Concept

Ignacio Belda^{1,2*}, Iratxe Zarraonaindia^{3,4}, Matthew Perisin¹, Antonio Palacios^{1,5} and Alberto Acedo^{1*}

¹ Biome Makers Inc., San Francisco, CA, USA, ² Department of Microbiology, Biology Faculty, Complutense University of Madrid, Madrid, Spain, ³ Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country, Leioa, Spain, ⁴ IKERBASQUE – Basque Foundation for Science, Bilbao, Spain, ⁵ Laboratorios Excell Iberica, Logroño, Spain

OPEN ACCESS

Edited by:

Sandra Torriani,
University of Verona, Italy

Reviewed by:

David Rodriguez-Lazaro,
University of Burgos, Spain
Braulio Esteve-Zarzoso,
Universitat Rovira i Virgili, Spain

*Correspondence:

Ignacio Belda
ignaciobel@ucm.es
Alberto Acedo
acedo@biomemakers.com

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 09 December 2016

Accepted: 21 April 2017

Published: 08 May 2017

Citation:

Belda I, Zarraonaindia I, Perisin M, Palacios A and Acedo A (2017) From Vineyard Soil to Wine Fermentation: Microbiome Approximations to Explain the “terroir” Concept.

Front. Microbiol. 8:821.
doi: 10.3389/fmicb.2017.00821

Wine originally emerged as a serendipitous mix of chemistry and biology, where microorganisms played a decisive role. From these ancient fermentations to the current monitored industrial processes, winegrowers and winemakers have been continuously changing their practices according to scientific knowledge and advances. A new enology direction is emerging and aiming to blend the complexity of spontaneous fermentations with industrial safety of monitored fermentations. In this context, wines with distinctive autochthonous peculiarities have a great acceptance among consumers, causing important economic returns. The concept of *terroir*, far from being a rural term, conceals a wide range of analytical parameters that are the basis of the knowledge-based enology trend. In this sense, the biological aspect of soils has been underestimated for years, when actually it contains a great microbial diversity. This soil-associated microbiota has been described as determinant, not only for the chemistry and nutritional properties of soils, but also for health, yield, and quality of the grapevine. Additionally, recent works describe the soil microbiome as the reservoir of the grapevine associated microbiota, and as a contributor to the final sensory properties of wines. To understand the crucial roles of microorganisms on the entire wine making process, we must understand their ecological niches, population dynamics, and relationships between ‘microbiome- vine health’ and ‘microbiome-wine metabolome.’ These are critical steps for designing precision enology practices. For that purpose, current metagenomic techniques are expanding from laboratories, to the food industry. This review focuses on the current knowledge about vine and wine microbiomes, with emphasis on their biological roles and the technical basis of next-generation sequencing pipelines. An overview of molecular and informatics tools is included and new directions are proposed, highlighting the importance of –omics technologies in wine research and industry.

Keywords: NGS, wine microbiome, vine health, soil microbiome, metagenomic analysis, bioinformatic tools and databases, 16S rRNA gene sequencing

INTRODUCTION

Wine is a product with high sociocultural interest. In particular, wines with distinctive autochthonous properties have a great demand among consumers and collectors, causing important economic consequences. It is well known that physical (climate) and biological factors (soil, grape variety and fauna), as well as viticulture and enological techniques work together to determine the sensory-characteristics of a wine from a particular region, establishing the concept of *terroir*. In this sense it should be noted that, apart from these factors, recent studies highlight the contribution of the native vine microbiota in the winemaking process of wines from a particular region (Knight et al., 2015; Bokulich et al., 2016). Additionally, results from Burns et al. (2016), Grangeteau et al. (2017) correlate human-agronomical practices in vineyards with the soil and grape microbiota and, also with its later behavior at cellar, reinforcing the interdependence between the anthropogenic and microbiological basis of *terroir*.

Microbes transform plant products into socio-economically important products and fermented beverages, such as wine, which is an extremely important sector for several countries. For instance, the International Organization of Wine and Vine (OIV) estimated in 2015 that the global wine-growing surface area was 7,534,000 hectares, with the biggest producer being Italy (18% of the global total), followed by France (17.3%) and Spain (13.5%). Outside the EU, the USA has the highest wine production followed by Argentina, Chile and Australia (OIV, 2015).

Due to the economic importance of the grapevine, this crop has received considerable interest among researchers; although this attention mainly focuses on the plant genome and transcriptome/metabolome to better understand how the plant responds to the physical environment, abiotic stresses and diseases (e.g., the International Grape Genome Program, IGGP). However, plants cannot be considered a self-contained, isolated organism, as plant fitness is a consequence of the plant *per se* and its associated microbiota (Vandenkoornhuys et al., 2015). Thus, a more holistic conception should include plant-microorganisms and microbe-microbe interactions.

Although the role of microorganisms at cellar stages has been well investigated, the biological aspect of soils has not received similar attention, when actually it contains a great microbial diversity with important roles in plant nutrition and health (Compan et al., 2010; Bhattacharyya and Jha, 2012). Next-generation sequencing (NGS) approaches have uncovered a higher than expected microbial diversity in both vine and wine and discovering new microbial species, some with unknown contributions to the organoleptic properties of wines (Bokulich et al., 2016). Stable differences among microbial populations of grape musts have been attributed to grape variety, geographical area, climatic factors and vine and grape health, leading to the concept of vine microbial *terroir* (Bokulich et al., 2014). This fact has been reinforced at a phenotype-metabolome level by other works such as Knight et al. (2015), Bokulich et al. (2016), and Belda et al. (2016). The later observed distinctive and clustered metabolic profiles (production of hydrolytic enzymes) for yeast strains depending on their geographical origin. It has been also observed that the origin of these microorganisms in musts is

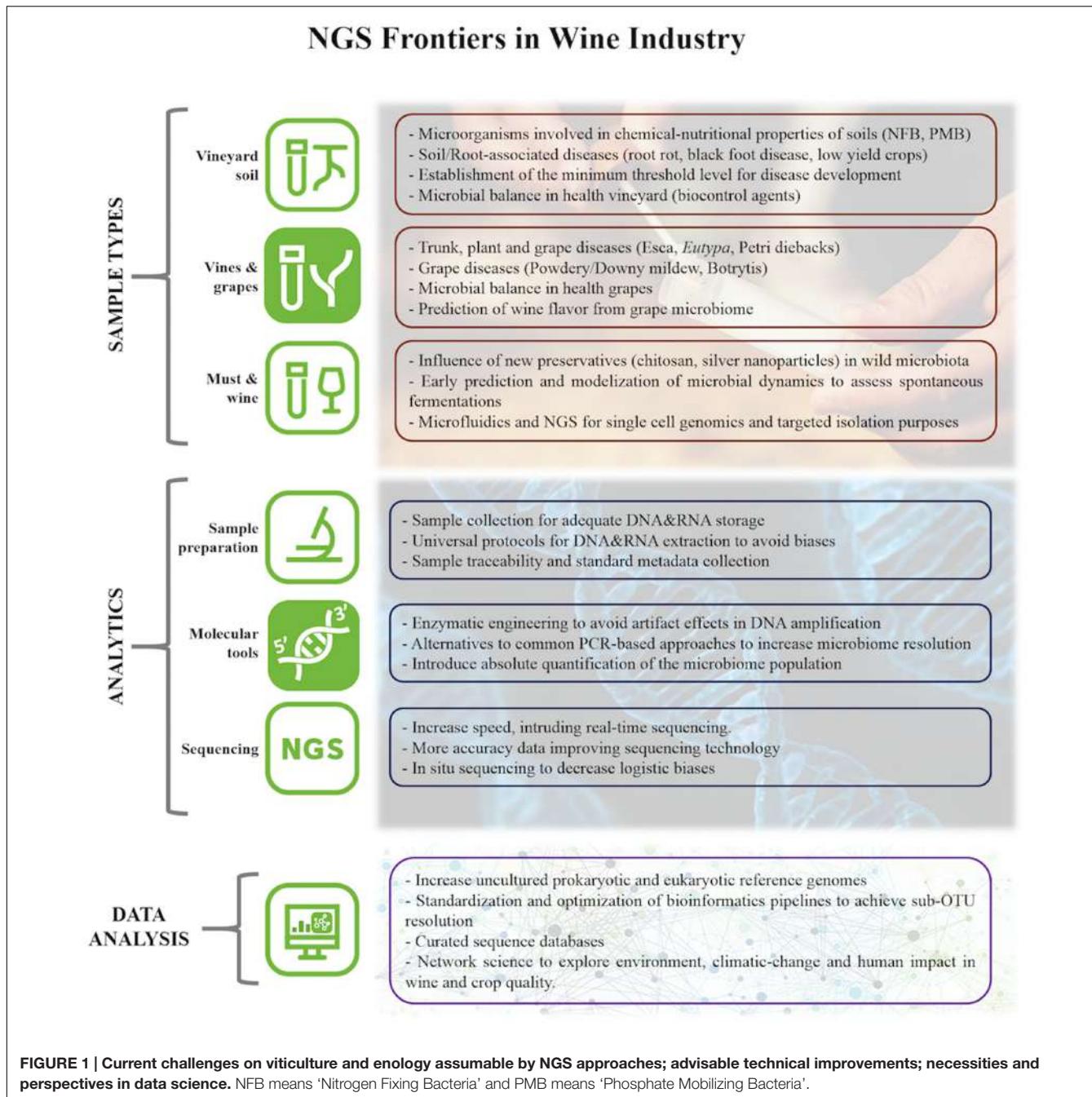
the microbial consortia of grapes, with the original reservoir of these microorganisms being vineyard soil (Zarraonaindia et al., 2015). Thus, the microbiological aspects of wine production are influenced by the vineyard and not just by the winery and fermentative processes.

The maturation of grapes is a complex process that depends on numerous factors (Kennedy, 2002). Traditionally, the most common measured parameters include: sugar concentration, acidity and aromatic and phenolic maturity. However, soil and grape microbiological complexity throughout the cycle of the vine and grape maturation is rarely taken into consideration.

Communities of microorganisms (fungi, yeast and bacteria) associated with the vineyard play an important role in soil productivity as well as disease resistance developed by the vine. It is important to understand the microbial consortia associated with particular diseases, such as *Esca*, *Eutypa*, *Botryosphaeria*, and *Phomopsis* diebacks, and also the dynamics of infection processes in order to take preventive actions, especially at the most critical moments (Figure 1). For instance, microbial insights are crucial for defining strategies for the preparation of new plantings. At this stage, it could be interesting to improve the microbiological conditions of the soil by bioremediation and to avoid risk of cross infection during pruning (Bertsch et al., 2013; Fontaine et al., 2016).

The diversity and number of microorganisms that are able to establish in an ecological niche in the soil and on the vine will determine both the grapes' health and the variability of microorganisms that will be introduced in the winery that further affect the fermentation processes and wine maturation (Barata et al., 2012). Thus, with adequately managed microbiome information, it could be possible to prevent fermentation problems, volatile acidity increases, *Brettanomyces* contamination and biogenic amines production. Knowing more about the microbiological conditions of the vineyard allows the winegrower to think about the reduction of chemical treatments and performing them only when they are objectively necessary. Additionally, this knowledge would help the winemaker to use lower sulfur concentration at cellar stages and even to decide the type of yeast and dose to be inoculated if and when necessary (Figure 1). This is valued information especially considering new enology trends, such as organic wines.

Next-generation sequencing technologies enable the detection and quantification of microorganisms present in vineyard soil, grapes, as well as its transformation later in winery. The impact of the microbiological component of *terroir* and how it contributes not only to its quality but also in the organoleptic features of the wine is considerable. This impact also contributes to the sensory regional distinctiveness and the wine style of the winery that currently plays an important role in differentiation and competitiveness in the worldwide market. If something can distinguish one vineyard from another, among other factors, it certainly is its microbial community. In this context, the objective of this review is to summarize the current knowledge about the role of microbial communities in viniculture, highlighting the contributions of NGS technologies and identifying new scientific-industrial frontiers.



THE MICROBIOME OF VINE AND WINE: A REVIEW

Plants host a variety of microorganisms (fungi, yeast, and bacteria) on and inside organs and their surrounding soil. Among these inhabitants are both harmful and beneficial microbes that are involved in crucial functions such as plant nutrition and plant resistance to biotic and abiotic stresses, hence in plant growth promotion, fruit yield, disease resistance and survival (Lugtenberg and Kamilova, 2009; Compat et al., 2010; Bhattacharyya and Jha, 2012).

Studies on microorganisms associated with grapevines have been centered on the cultivable fungi (mainly yeast) or bacteria that can have a negative economic impact, compromising the yield and quality of the grapevine, as well as wine production. Studies have focused on disease causing pathogens (*Agrobacterium vitis*, *Xylella fastidiosa*, *Erysiphe necator*, *Phomopsis viticola*, *Fusarium* spp., etc.) and microorganisms of enological interest. The later species have been grouped into three classes [reviewed in Barata et al. (2012)]: (1) easily controllable or innocent species, without the ability to spoil wine when good manufacturing practices are applied; (2) fermenting species

responsible for sugar and malic acid conversion; and (3) spoilage *sensu stricto* species responsible for wine alteration. The most widely known cultivable bacteria are acetic acid bacteria (AAB; e.g., *Acetobacter* and *Gluconacetobacter*) and lactic acid bacteria (LAB; e.g., *Lactobacillus*, *Oenococcus*, and *Pediococcus*). Among yeasts, *Saccharomyces* members have attracted most of the attention as they are the main fermentation agents commonly used as inocula (e.g., *Saccharomyces cerevisiae*, *S. bayanus*, *S. pastorianus*, and *S. paradoxus* among others), while other genera are the most frequent wine spoilers (e.g., *Brettanomyces/Dekkera*, *Issatchenkia*, *Zygoascus*, and *Zygosaccharomyces*).

While culture dependent methods have been useful to detect and identify microbial organisms associated with grapevine and grape products, and also to study *in vitro* their metabolic properties (Belda et al., 2016), they have led to a rather biased picture of the microbial community. These methods neglect the larger, non-culturable fraction that is believed to be as high as the 95–99% of the microorganisms present (Amann et al., 1995; Curtis, 2002). In wine environment, due to the stressful environment associated to the addition of SO₂, high ethanol concentration, etc., a fraction of the bacteria and yeast enter in a Viable But Non-Culturable state (VBNC) (Millet and Lonvaud-Funel, 2000; Divol and Lonvaud-Funel, 2005). At this state cells do not grow on culture media, however, they are still viable and maintain a detectable metabolic activity (Yamamoto, 2000) which may affect fermentation performance as well as flavor. Examples of such microorganisms include *Candida stellata*, *Brettanomyces bruxellensis*, *S. cerevisiae*, *Zygosaccharomyces bailii*, etc. (Salma et al., 2013). Thus, in order to reach to these VBNC microbiologists were driven to develop alternative culture-independent techniques. Particularly, quantitative real time PCR (qPCR) has been widely used to detect bacteria and yeast considered to be wine spoilers and that have VBNC strains responsible for the production of off-flavors or having a negative impact on wine, e.g., *Brettanomyces* spp. (Tofalo et al., 2012). Nowadays, qPCR is believed to be a rapid diagnostic tool to detect the presence and quantify the abundance of particular microorganisms of interest, however, when the objective is not a targeted species, but rather a whole community analysis, PCR-DGGE has been the classical method of choice. The later technique is adequate to approximate the total community profile and for comparative community structure analysis, but it has several drawbacks mainly associated to biases related with species richness estimates and its low sensitivity to detect low abundance species (Neilson et al., 2013). For instance, multiple bands could associate with single isolates. In addition, multiple sequences might be associated with a single band and preferential amplification biases between phylogenetically diverse members of the community have been shown (Neilson et al., 2013). Andorrà et al. (2010) compared the population dynamics of microorganisms of grape must fermentation by three culture independent techniques (DGGE, direct cloning of amplified DNA, and qPCR) with plate counting, and evidenced that the biodiversity observed in the must and at the beginning of fermentation was much higher when DGGE or direct cloning were used. However, the predominance of certain yeast such as

C. zemplinina and *S. cerevisiae* during fermentation limited the detection of low abundant species. Thus, while DGGE is believed to give a quick and non-expensive view of the community, it skews microbial diversity estimates (David et al., 2014) and it has a limited use to study diverse environmental samples dominated by few species (Andorrà et al., 2010). When adding NGS technique into the detectability comparition of culture independent techniques to study yeast community in must and ferments, the studies evidenced that larger numbers of yeast species were detectable by NGS than by PCR-ITS-RFLP or DGGE in grape samples. Moreover NGS detected species in ferment samples that were undetectable with the two later techniques (David et al., 2014). In addition, Wang et al. (2015) analyzed Carignan and Granache grape must and fermentation from three vineyards in Priorat (Spain) and found that NGS detected all the species identified by the rest of methods (DGGE, qPCR and culture dependent), whereas DGGE could just detect the dominant species of non-*Saccharomycetes* class. Thus, NGS showed to be more appropriate to understand must and wine environment yeast communities (David et al., 2014; Wang et al., 2015).

Next-generation sequencing technologies are providing a powerful approach to achieve a more complete understanding of the complexities of microbial communities and their impact on plant growth, disease resistance/susceptibility, climate adaptation and environmental remediation. This technology is enabling researchers to simultaneously obtain information on thousands of taxa as opposed to targeted approaches that detect only a taxonomically predefined group. Thus, metagenomics coupled with new bioinformatics tools, is allowing performance of more complex multifactorial analyses and is becoming a powerful strategy in diagnostics, monitoring, and traceability of products. Its application in viticulture while recent is promising (Table 1), as accumulating data suggest that there is a much higher microbial diversity associated both with the plant (Leveau and Tech, 2010; Pinto et al., 2014; Zarraonaindia et al., 2015) and the fermentation process (Bokulich et al., 2012; Piao et al., 2015; Pinto et al., 2015; Portillo and Mas, 2016; Stefanini et al., 2016) compared to previous culture based studies. Most metagenomics research in this field has focused on microbial monitoring during fermentation to obtain a detailed description of the relevant microbial populations associated with grape and must that might lead to wine spoilage, an advance highly valuable for winemaking. These NGS-enabled studies reflect a wider range of bacteria, besides the commonly detected LAB and acetic acid species, able to persist in fermenting musts of various grape varieties (Bokulich et al., 2012; Piao et al., 2015; Portillo and Mas, 2016; Stefanini et al., 2016). For instance, the first wine-related study conducted in the wine environment with NGS was conducted by Bokulich et al. (2012) during botrytized wine fermentation using 16S rRNA gene amplicon sequencing. These authors showed an array of fluctuating low abundant taxa not traditionally associated with wine, as well as atypical LAB communities during the process. Similarly, results from Portillo and Mas (2016) suggested that AAB are more abundant and dynamic than previously thought during low or unsulfited wine fermentations, and seemed to be independent of the grape variety. Interestingly, in this

TABLE 1 | Research and industrial hallmarks of viticulture and enology led by NGS approaches.

Ecological features addressed by microbiome study	Reference
Interference of microorganisms in plant physiology	Lugtenberg and Kamilova, 2009; Compart et al., 2010; Bhattacharyya and Jha, 2012; Martins et al., 2013; Vandenkoornhuyse et al., 2015
Microbial diversity in vineyard	Leveau and Tech, 2010; Pinto et al., 2014; Zarraonaindia et al., 2015
Microbial diversity in wine fermentations	Bokulich et al., 2012; Piao et al., 2015; Pinto et al., 2015; Portillo and Mas, 2016; Stefanini et al., 2016
Anthropogenic-agronomical practices determining vineyard and wine microbiota	Burns et al., 2016; Graneteau et al., 2017
Microbial contribution to wine chemistry	Verginer et al., 2010; Bokulich et al., 2016
Terroir markers (microbial zoning)	Bokulich et al., 2014, 2016; Burns et al., 2015; Knight et al., 2015

study yeast diversity and dynamics during wine fermentation was assessed in addition to bacteria, evidencing that the genera *Hanseniaspora* and *Candida* were dominant during the initial and mid-spontaneous fermentation of Grenache grapes while certain *Candida* and *Saccharomyces* species predominated at the end of the fermentation. Other studies have demonstrated how different fermentation techniques (spontaneous vs. inoculated) affect the microbial community composition and its succession during fermentation (Piao et al., 2015), and also how the previous agronomical practices in the vineyard could play a critical role in these population dynamics (Graneteau et al., 2017). These authors' results indicated certain phyla are associated with each particular technique. Interestingly, they observed that *Gluconobacter* experienced a notable increase during organic fermentation, which led the authors to conclude that this might explain the increased susceptibility to wine spoilage in wines produced using that technique.

These above-mentioned studies enhance our understanding of microbial diversity during fermentation and allow the identification microbial contamination sources. However, as DNA sequencing approaches detects living as well as dead microorganisms, it is still not clear to what extent these microorganisms metabolically are active and capable of affecting organoleptic properties of wine. The role of the microbiota influencing the flavor, color and quality of wine, under a systems biology perspective, remained elusive until recently.

While soil, weather, farming techniques and grape variety contribute to the unique qualities of wine, adding distinctiveness and thus market value, the contribution of the microbiota in defining *terroir* is now in the spotlight of scientific research. Regionally distinct wines are highly appreciated by consumers and add value to the industry. In Spain alone there are

90 zones, which produce distinct so-called PDO wines, of which 69 are Denomination of Origin (DO), 2 are Qualified Denomination of Origin (DOCa), 7 are Quality Wine with a Geographical Indication (Vino de Calidad) and 14 are Single Estate Wine (Vino de Pago). While the chemosensory distinction of wines from different growing regions has been previously established [e.g., Loópez-Riterto et al. (2012)], indigenous microorganisms associated with grapes were shown to be able to produce compounds responsible for the regional flavors of the resulting wine, e.g., VOCs (Verginer et al., 2010). In addition, Knight et al. (2015) experimentally demonstrated that wine organoleptic characteristics are affected by the origin and genetics of wild *S. cerevisiae* natural strains, providing objective evidence for a microbial aspect to *terroir*. Bokulich et al. (2014) showed that Cabernet Sauvignon must from different growing regions in California could be distinguished based on the abundance of several key fungal and bacterial taxa. This differential must microbiota could potentially influence wine properties and contribute to the regionalization of wine. The later was further proved in Bokulich et al. (2016); these authors demonstrated that both grape microbiota and wine metabolite profiles were able to distinguish viticultural area designations and individual vineyards within Napa and Sonoma Counties in CA, USA. Interestingly, the vineyard microbiota correlated with the chemical composition of the finished wines, hinting at the possibility of predicting wine phenotypes prior to fermentation. Nevertheless, wine aroma is defined by hundreds of chemical compounds with different natures (i.e., higher alcohols, esters, fatty acids, terpenes, thiols) causing a broad spectrum of sensory thresholds, and also suffering synergies and antagonisms (Belda et al., 2017). Thus, looking for microbial signatures determining wine typicity, the sensorial characterization of wines should consider not only chromatographic analysis (revealing the diversity and concentration of aroma compounds), but also developing serious sensorial or olfactometry analysis to reflect the real perception of wine aroma or, at least, considering odor activity values (OAVs) to correlate the real influence of microbial species in wine aroma, as was addressed by Knight et al. (2015).

While grape and must have been more heavily researched, Zarraonaindia et al. (2015) further hypothesized that the soil and its associated microbiota influences wine characteristics. First, per these authors' studies, the aboveground bacterial community was significantly influenced by soil edaphic factors such as total carbon, moisture and soil temperature, which would ultimately impact the quality of grapes due to changes in nutrient availability for the plant. Second, soil bacterial communities differed between the sampled vineyards in Long Island, New York and those differences were reflected in the microbial composition in vine roots. These root endophytes can shape the microbial assemblages of aboveground organs by changing the endophytic microbial loads in grapes. Third, a significant input of soil microorganisms to grapes through epiphytic migration during harvest was suggested. The later was also evidenced by Martins et al. (2013), leading Zarraonaindia et al. (2015) to propose that soil derived microorganisms could have a greater role than previously anticipated in wine, as they will ultimately end up in the fermentation tanks. The link between soil microbiota

and *terroir* was further evidenced by Burns et al. (2015) who identified distinctive microbial community profiles by American Viticultural Areas (AVA).

NGS MICROBIAL PROFILING: KEY STEPS, BIASES AND LIMITATIONS

The above summarized studies were conducted on grapevines and wine and address microbial composition by means of 16s rDNA PCR amplicon and ITS (Internal Transcribed spacer) NGS sequencing for elucidating the bacterial and fungi community, respectively. This marker gene amplification and sequencing method, also called amplicon sequencing, has become the method of choice to simultaneously detect multiple species in must and wine environment since 2012 (see Bokulich et al., 2012, 2016; Pinto et al., 2014, 2015; Knight et al., 2015, among others). However, the particular experimental question of the research to be conducted will determine the method mostly suited to answer to the question. For instance, if the goal is to track a particular microbial strain or genus from soil to must to fermentation, then qPCR could be a more appropriate and has the added benefit of absolute quantification (Neely et al., 2005). To detect a specific microbe, primers must be designed to be highly specific for the microbe of interest. Often the primer design can be completed by genome comparison of targeted and non-targeted strains to find a unique gene or region. Another strategy involves targeting a conserved gene (16S rRNA, *gyrB*, *rpoB*) and making sure the primers mismatch off-target strains particularly at the 3' end. Single copy genes provide an added bonus for absolute quantification. Microbe quantification by qPCR, however, does not scale easily if the goal is to analyze more than a few strains while amplicon sequencing is suited to determine the community.

However, amplicon sequencing is not free of pitfalls, and different biases have been described in multiple steps of the process; First, DNA extraction method is one of the key and limiting steps for metagenomic analysis by NGS. Various approaches have been applied for environmental DNA extraction, including freeze-thaw lysis (Herrick et al., 1993), bead beating (Miller et al., 1999; Courtois et al., 2001; Urakawa et al., 2010; Petric et al., 2011), liquid nitrogen grinding (Ranjard et al., 1998), ultrasonication (Picard et al., 1992), hot detergent treatment (Holben, 1994), use of strong chaotropic agents like guanidinium salts (Porteous et al., 1997), and high concentration of lysozyme treatment (Hilger and Myrold, 1991). Furthermore, soil, grapes and wine are complex physicochemical environmental samples that contain many interfering agents for molecular analysis such as impurities, phenols, humic acid, fulvic acid, metal ions and salts, and therefore additional purification steps are necessary which can introduce bias by altering the original community (e.g., a fraction of the community might be lost through purification, etc.). There are several commercial kits that could be used to fasten the process, however, the selection of the best DNA extraction method and kit is not straightforward as different DNA extraction methods can produce different results (Keisam et al., 2016). Unfortunately,

there is no “gold standard” for DNA extraction method and one should be selected on a case-by-case basis considering the aims, specimens of the study and scalability (including simplicity, cost effectiveness, and short handling time) and intended study comparisons. An additional problem is the introduction of contaminating microbial DNA during sample preparation. Possible sources of DNA contamination include molecular biology grade water, PCR reagents and DNA extraction kits themselves. Contaminating sequences matching water-and soil-associated bacterial genera including *Acinetobacter*, *Alcaligenes*, *Bacillus*, *Bradyrhizobium*, *Herbaspirillum*, *Legionella*, *Leifsonia*, *Mesorhizobium*, *Methylobacterium*, *Microbacterium*, *Novosphingobium*, *Pseudomonas*, *Ralstonia*, *Sphingomonas*, *Stenotrophomonas*, and *Xanthomonas* have been reported previously. The presence of contaminating DNA is a particular challenge for researchers working with samples containing a low microbial biomass. In these cases, the low amount of starting material may be effectively swamped by the contaminating DNA and generate misleading results (Salter et al., 2014).

Second, DNA library preparation, based on fragment amplification through PCR with barcoded primers, is another step in which it is possible to introduce additional biases. The choice of primers and targeted variable regions will bias identification and quantification (Soergel et al., 2012; Bokulich and Mills, 2013). Additionally, in any PCR- and primer-based taxonomic investigation, members of a microbial community may be omitted, distorted, and/or misrepresented, typically due to primer mismatches or PCR biases (Acinas et al., 2005; Hong et al., 2009; Lee et al., 2012; Pinto and Raskin, 2012; Logares et al., 2014). On the contrary, primers might show variability in their amplification efficiency by for example, favoring certain species amplification (Baker et al., 2003; Sipos et al., 2007; Klindworth et al., 2013). This preferential amplification is thought to be derived from different sources such as primer mismatches, the annealing temperature and PCR cycle numbers (Sipos et al., 2007). For instance, Sipos et al. (2007)' studies evidenced that *A. hydrophila* and *P. fluorescens* were preferentially amplified over both *Bacillus* strains when the 63F primer was used (which contained three mismatches against DNA isolated from the *Bacillus* strains), while the 27F primer amplified all templates without bias. Interestingly, the bias introduced by primer mismatches was reduced at lower annealing temperatures.

Multiple primer pairs are available for marker genes, and each pair is associated with its own taxon biases. Marker gene databases are frequently updated, and the updated information can include new microbial lineages with suboptimal or poor binding to existing PCR primers; to maximize taxonomic sensitivity in light of these new data, primers may need to be periodically redesigned. A recent example in the literature is the modification of the most common 16S primers used 515f and 806r to remove known biases against *Crenarchaeota/Thaumarchaeota* and the marine and freshwater Alphaproteobacterial clade SAR11 (Apprill et al., 2015; Parada et al., 2016).

Klindworth et al. (2013) evaluated the coverage and phylum spectrum for bacteria and archaea of 175 primers and

512 primer pairs *in silico* for three amplicon size classes (100–400, 400–1000, >1000 bp), demonstrating the differences in coverage and specificity among the studied primers. Besides, this information represents a valuable guideline for selecting primer pairs that could minimize the bias in PCR-based microbial diversity studies. In the same way, probeBase¹ is an additional online resource, providing the opportunity to evaluate the *in silico* hybridization performance of oligonucleotides, as well as finding suitable hierarchical probes that could target an organism or taxon of interest at different taxonomic levels (Greuter et al., 2016).

The ideal marker gene should have conserved regions that flank variable regions. The conserved regions allow primer design to amplify multiple taxons at ones. Ribosomal rRNA genes fit this description and have been widely used for identification of bacteria/archaea (16S) and fungi (ITS) (Gilbert et al., 2010). However, ribosomal RNA genes show copy-number variation, with very disparate number of copies per taxa (from one in many species to up to 15 in some bacteria and to hundreds in some microbial eukaryotes) biasing conclusions related to the abundance of the organisms.

To evaluate the entire microbial community in the specific case of the wine ecosystem, it is necessary to strike an appropriate balance between amplifying all members of every taxon (high coverage) and obtaining the highest taxonomic resolution possible, e.g., to be able to discriminate among closely related species (**Figure 1**). Each marker shows differences in its discrimination power at intra-genera as well as at intra-species level. Thus researchers must have that in mind when designing their project, in order to choose the most appropriate molecular marker to answer their particular question/s. For instance, primer pair 515f/806r is the most widely used for targeting the V4 region of for bacteria/archaea (Parada et al., 2016), and this combined with Illumina sequencing has been used to characterize the microbiomes of numerous environments (Caporaso et al., 2012), vine and wine environments among them. Data from high diverse environments, as Sakinaw Lake, showed species resolution level from 49.4% of the 16S V4 sequences classified compare with 74.5% using full 16S. Although the relative classification differences at the sequence level do not directly translate to differences in community representation (Singer et al., 2016). However, vine and wine samples have the added difficulty in that mitochondrial and chloroplast DNA can be amplified with these V4 region primers and thus grapevine plastid sequences overwhelm the sequencing. Researcher have two ways to avoid this problem: design primers that mismatch mitochondrial/chloroplast sequences or add blocking reagents that bind these sequences (Lundberg et al., 2013). Besides, the V4 domain of the 16s rDNA gene is considered to be the most suitable marker for capturing the bacterial community in wine, as it is able to reliably discriminate LAB to genus-level (Bokulich et al., 2012). However, in fermentative systems, some species of LAB are considered wine spoilers while others exhibit malolactic activity, thus it might be essential to reach to species level (Bokulich and Mills, 2012) and/or strain level, in order to have

a more comprehensive view of the community. Unfortunately, currently available amplicon sequencing markers are unable to capture that level of resolution in all taxa. These limitations could be overcome by combining several techniques such as genera specific T-RFLP or qPCR and amplicon sequencing.

Third, important sources of artifacts are also derived from the High-throughput sequencing technology chosen. While pyrosequencing introduces homopolymer errors (indel error), Illumina sequencing has average substitution errors at 0,0086 sequencing rate (Schirmer et al., 2015). Sequencing platforms also show a disparity in sequencing depth (number of reads per run) and read length. Illumina MiSeq is the most commonly used sequencer for amplicon sequencing due to its high coverage with a total nucleotide sequenced of 15GB allowing sequencing the abundant and rare community giving a deep view of the community composition. However, Illumina sequencing is characterized by a short variable region sequencing (2 × 300 bp vs. 700 bp in 454). Currently, nearly full-length rRNA gene sequencing is possible with PacBio and Nanopore technologies (Benitez-Paez et al., 2016; Schloss et al., 2016).

Finally, one of the biggest limitation of amplicon sequencing techniques relays on its inability to address a functional characterization of the microbial communities. There are many desired microbial functions in winemaking, mainly related to alcoholic and malolactic fermentations, and diversity of genes related to those functions may influence winemaking more than just taxonomic diversity. In addition, closely related strains with highly similar 16S rRNA gene or ITS sequences contain different fermentation-related genes (Knight et al., 2015) and thus that strain diversity remains hidden in current amplicon sequencing studies. Single-cell genomics emerges as a potential strategy that could help to obtain a deeper knowledge into species-strain level diversity. This strategy is powerful when the targeted organism is dominant or high abundant in a low species richness ecosystem. However, in highly diverse ecosystems or when the species to be targeted is low abundance, it may require a higher sorting throughput, specific labeling with fluorescent probes or a previous cultivation step, all of which could contribute to biases.

Alternatively, shotgun metagenomic sequencing would also reveal functional genes in addition to rRNA genes, allowing a more comprehensive genomic and functional representation through whole-genome sequencing (WGS) of complete communities, but the cost and the number of reads needed to estimate the environmental population is high compared to PCR-based approaches. Even more in wine samples, as a very deep sequencing is required to detect microbes due to an overabundance of plant DNA (Zarraonaindia et al., 2015), making this method costly for a large number of samples.

Metatranscriptomics is emerging as a powerful technology for the functional characterization of microbial communities that can reveal both the taxonomic composition and active biochemical functions of the detected organisms. These approach is of especial interest in wine environment, as amplicon sequencing is not able to discriminate among living or dead organisms, nor the metabolically active or inactive organisms. However, the high sequencing depth needed and the high cost associated with the sequencing of each sample limits the number

¹<http://www.probebase.net>

of samples that could be surveyed within a project currently. In addition, challenges associated with this technique include among others, the lack of established reference genomes to annotate the short reads generated in the sequencing and the high computational effort needed for the analyzes. Being a technique still in its infancy, new analysis tools and standardized pipelines are under development. In this context, the next section aims to summarize critical concepts and sources of biases in NGS analysis.

BIOINFORMATICS AND PREDICTIVE METHODS TO UNCOVER THE MICROBIAL TERROIR

Along with the relative ease with which thousands of organisms can be detected in samples via 16S/ITS sequencing, a whole host of bioinformatics approaches have been developed to extract meaningful results from the large datasets that are generated. The bioinformatics challenge comes in at least two parts (I) preprocessing the datasets into a collection of representative reads (or operational taxonomic unit – OTU) that can be associated with databases of known species and (II) associating the collection of species inferred in a sample (known and newly detected) with properties of the sample in order to study the relationships between the microbiome and the *terroir*.

In the first stage, the large amounts of raw sequencing reads are processed (trimming adaptor sequences, merging forward and reverse sequences, filtering on read quality) before finally being dereplicated into a collection of unique sequences. There is a lot of software available to perform these tasks and they often are part of packages that offer an entire processing pipeline (USEARCH, vsearch, FASTX-Toolkit) (Edgar, 2013; Rognes et al., 2016). The unique sequences are then clustered according to sequence similarity, choosing a relatively arbitrary cutoff at 97% identity (Seguritan and Rohwer, 2001), resulting in a set of OTUs that are each assumed to be originating from a specific organism. In other words, OTUs are proxies for microbial species in the sample (Schloss et al., 2009; Caporaso et al., 2010).

Although conceptually simple, this step poses major challenges both computationally and in terms of biases that might potentially bleed into subsequent analysis. First of all, for large sets of sequences, all against all pairwise alignments would be prohibitive, e.g., 1 million of unique sequences (commonly encountered), would require 1000 billion pairwise comparisons. This has led to comprehensive bioinformatics pipelines for OTU clustering, including the software pipelines mentioned above (USEARCH, vsearch, swarm), which all rely on clever heuristics (Edgar, 2013; Eren et al., 2013; Mahé et al., 2014, 2015; Tikhonov et al., 2015; Rognes et al., 2016) in order to accelerate this process at the expense of perfectly accurate clustering. The second challenge is to avoid biases that can occur during OTU clustering. The biases can be multifold; (a) different biological species might have the same sequence and therefore be grouped into one set, (b) sequencing errors or amplification errors (including chimeric reads) or untrimmed sequences can group sequences that have the same origin into separate groups. The

first issue will underestimate biological diversity whereas the latter will overestimate it. Together these scenarios will corrupt the accurate representation of the real biological makeup of the *terroir*. This highlight again that it is important to quality trim and filter the raw sequences to minimize the risk of including artifacts in environmental data sets.

Finally, the curated OTUs are subjected to phylogenetic assignment, which aims to identify what species or genus an OTU most likely belongs to. This is achieved by comparing them with taxonomically classified sequences at databases, such as GreenGenes (for bacteria community characterization), SILVA (bacteria and eucaryotes) and Unite (for Fungi) among others. Again, a range of software is available (Qiime, UTAX, SINTAX, stampa) (Caporaso et al., 2010; Edgar, 2013, 2016). This stage is again a source of biases, partly because OTUs can represent multiple species, there is ambiguous assignment, and because too small differences that do exist could be ignored by these methods. For instance, in the case of Oligotyping, a single base pair can differentiate ecological strains (Eren et al., 2013). Furthermore, and more generally, reference databases are themselves based largely on predicted species rather than experimentally cultivated species and can thus bias taxonomic assignment. Additionally, different reference databases would yield different taxonomic assignments as a function of completeness and quality of the database (McDonald et al., 2012). Notably, if a given species is not represented within the database, sequences derived from that species would receive an incorrect assignment or remain unclassified. This is aggravated for wine and soil associated microbial sequences field, where reference databases lag behind human-associated microbes. Increasing and curating robust databases is a key goal for the scientific community (Figure 1). There are also other methods allowing comparison of amplicons derived from functional genes in which we might not know percent identities that correspond to taxonomic levels, but in some cases, are optimal to reflect geographical (and thus, environmental) distance (Haggerty and Dinsdale, 2017). In relation to wine related samples, cluster free methods show the potential to define the microbial *terroir* at the strain or sub-OTU level (Tikhonov et al., 2015; Eren et al., 2016).

Equipped with a dataset of biological entities in the *terroir* (genus- or species-level), the second bioinformatics challenge concerns associating the microbiome to the properties of the *terroir*. Depending on the aim, this can be more or less difficult. One goal is to use microbial community data to classify soil samples into types and geographical regions and, therefore, define the microbial *terroir*. Recently, Bokulich et al. (2016) demonstrated the power of this approach for classifying Californian regions and fermentation metabolites based on microbial abundances in musts. However, if species or even strain information is required to establish an association between microbiome and specific wine making properties, then the taxonomic assignment is essential and can make or break an analysis depending on the resolution it achieves and the biases it can prevent.

Apart from the nature of the question, generally, the structure of OTU abundance data poses some challenges that need to be carefully taken into account. Because the species can occur

in very different abundances (often spanning several orders of magnitude), the collection of species across samples can greatly vary. This leads to a very sparse dataset, which is defined as a dataset with many zero values. These zero values can be problematic as they could entertain multiple hypotheses; for instance, a zero count in a sample could be because a species is not present, or because it just has not been detected. This can lead to biased comparisons between samples. One way to deal with this is to use distance metrics that do not consider these situation (e.g., Bray-Curtis) or that specifically include a phylogenetic tree that allows to relate species information into meaningful groups. Preprocessing of OTU data from raw counts to a value that makes samples comparable to each other is the next step. This is also referred to as normalization and there are number of analytical choices available (Segata et al., 2011; Paulson et al., 2013) depending on whether low-abundance species or high abundance species should have more of an impact in the analysis in question. For instance, counts can be converted to frequencies (divide the number of reads by the total number of reads in the sample). The performance of these techniques given OTU table peculiarities has been tested elsewhere (McMurdie and Holmes, 2014; Weiss et al., 2016). This is also a crucial step when applying machine learning techniques.

With a preprocessed dataset available, the probe community level differences between samples, can be studied with supervised and unsupervised machine learning techniques. Unsupervised learning categorizes samples based on OTU abundances without prior knowledge of the sample phenotypes. Principle component analysis (PCA or more commonly PCoA) and clustering algorithms can be used to gain a high level view over differences in samples. These analyses are largely exploratory and provide visual evidence of community differences. If information regarding the *terroir* is available, or there are some clearly defined groups that are to be studied, supervised learning techniques can be applied to for instance classify new samples based on past community characterizations (Bokulich et al., 2016). Distinct wine regions, types and tastes make wine related samples well-suited for these classification methods (Statnikov et al., 2013). Different software packages are available to perform these methods and can be more or less adapted to the study of metagenomics problems (vegan, phyloseq, Qiime, mothur).

Another method to extract knowledge from microbiome data is to consider it as a network of interactions between individual strains. Aside from the impact of single strains in plant health (pathogens, symbionts) and wine characteristics, or spoilage potential, these strains impact wine production not in isolation but instead as members of complex microbial communities. Much research now focuses on these community level effects that can impact plant phenotypes such as flowering time (Wagner et al., 2014).

These predictive technologies allow to make initial inferences about whether these differentially abundant single OTUs cause certain phenotypes. However, they will require further testing, likely with pure culture treatments. One excellent example of going from correlation to causation is the use of pure

fungal and oomycete cultures in a common garden to confirm single strain effects on the overall microbial community structure associated with *Arabidopsis thaliana* (Agler et al., 2016).

Defining the microbial *terroir* with bioinformatics is only an early step to understand how microbes shape each step in winemaking. Wine imparts its taste and smell via metabolites, many derived from the grapes and many derived from or modified by microbes. Identifying which microbes influence these processes is key to defining how they affect the sensory profile of wines. As we add genomic sequences to our reference database we will be able to leverage annotated sequences to predict metabolic capacity for each microbe. Genome scale metabolic models (GSMM) combined with flux balance analysis allows for analysis of metabolic outputs given a set of inputs (Varma and Palsson, 1994). Furthermore, GSMMs can expand to community-level models (Zomorrodi and Maranas, 2012; Khandelwal et al., 2013; Louca and Doebeli, 2015) to uncover how microbes synergistically create complex wine metabolite profiles. Going forward, it will be critical not only to define which microbes created your favorite wine but also how their metabolisms shaped the taste of that wine. Thus, viticulture will benefit very much from the generation of commercial platforms that enable studying vine and wine microbiome and wine metabolome. Currently such platforms are already in place, with WineSeq[®] – (Biome Makers, Inc.)² allowing wine microbiome characterization through NGS and Wine Screener[®] – (Bruker)³ allowing wine metabolome analysis by nuclear magnetic resonance. These tools are based on robust databases and allow both producers and regulatory councils from appellations of origins to establish ‘standard profiles’ for their wines, and better understand the microbial and chemical bases of their distinctive *terroir*.

CONCLUSION

In this article, the impact of NGS technologies in vine and wine microbiology has been reviewed. Regarding the importance of microbiome in viticulture and enology, the role of microorganisms in the chemical and nutritional properties of vineyard soils, crop health and yield, and also in the later fermentation performance and wine flavor are the main challenges to explore using -omics tools. For that purpose, certain technical aspects should be improved at laboratory stages, such as universal DNA&RNA extraction protocols to avoid biases, and improved sequencing approaches to increase microbiome resolution and quantification. It is also important to develop robust and curated databases to improve taxonomic assignments (Figure 1). Finally, it is time to develop big data works, using statistical data-mining and machine learning tools to solve, in a holistic systems-biology view, the above-mentioned challenges in wine industry.

²<http://www.wineseq.com/>

³<http://www.bruker.com/products/mr/nmr/food-screener/wine-profiling/overview.html>

AUTHOR CONTRIBUTIONS

IB and AA conceived the work. IB, AP, and AA wrote the “Introduction” section. IB and IZ wrote “The microbiome of vine and wine: a review” section. AA, IZ, and MP wrote “NGS microbial profiling: key steps, biases and limitations” section. IZ and MP wrote “Bioinformatics and predictive methods to

uncover the microbial terroir” section. Finally, IB edited the final version of the manuscript.

FUNDING

This study was funded by WineSeq Project, BiomeMakers Inc.

REFERENCES

- Acinas, S. G., Sarma-Rupavartam, R., Klepac-Ceraj, V., and Polz, M. F. (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol.* 71, 8966–8969. doi: 10.1128/AEM.71.12.8966-8969.2005
- Agler, M. T., Ruhe, J., Kroll, S., Morhenn, C., Kim, S. T., Weigel, D., et al. (2016). Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biol.* 14:e1002352. doi: 10.1371/journal.pbio.1002352
- Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* 59, 143–169.
- Andorrà, I., Landi, S., Mas, A., Esteve-Zarzoso, B., and Guillamón, J. M. (2010). Effect of fermentation temperature on microbial population evolution using culture-independent and dependent techniques. *Food Res. Int.* 43, 773–779. doi: 10.1016/j.foodres.2009.11.014
- Apprill, A., McNally, S., Parsons, R., and Weber, L. (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.* 75, 129–137. doi: 10.3354/ame01753
- Baker, G. C., Smith, J. J., and Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* 55, 541–555. doi: 10.1016/j.mimet.2003.08.009
- Barata, A., Malfeito-Ferreira, M., and Loureiro, V. (2012). The microbial ecology of wine grape berries. *Int. J. Food Microbiol.* 153, 243–259. doi: 10.1016/j.ijfoodmicro.2011.11.025
- Belda, I., Ruiz, J., Alastruey-Izquierdo, A., Navascués, E., Marquina, D., and Santos, A. (2016). Unraveling the enzymatic basis of wine “flavorome”: a phylo-functional study of wine related yeast species. *Front. Microbiol.* 7:12. doi: 10.3389/fmicb.2016.00012
- Belda, I., Ruiz, J., Esteban-Fernández, A., Navascués, E., Marquina, D., Santos, A., et al. (2017). Microbial contribution to wine aroma and its intended use for wine quality improvement. *Molecules* 22, E189. doi: 10.3390/molecules22020189
- Benitez-Paez, A., Portune, K. J., and Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION portable nanopore sequencer. *Gigascience* 5, 4. doi: 10.1186/s13742-016-0111-z
- Bertsch, C., Ramírez-Suero, M., Magnin-Robert, M., Larignon, P., Chong, J., Abou-Mansour, E., et al. (2013). Grapevine trunk diseases: complex and still poorly understood. *Plant Pathol.* 62, 243–265. doi: 10.1111/j.1365-3059.2012.02674.x
- Bhattacharyya, P., and Jha, D. (2012). Plant growth-promoting rhizobacteria (PGPR): emergence in agriculture. *World J. Microb. Biot.* 28, 1327–1350. doi: 10.1007/s11274-011-0979-9
- Bokulich, N. A., Collins, T. S., Masarweh, C., Allen, G., Heymann, H., Ebeler, S. E., et al. (2016). Associations among wine grape microbiome, metabolome, and fermentation behavior suggest microbial contribution to regional wine characteristics. *mBio* 7, e00631-16. doi: 10.1128/mBio.00631-16
- Bokulich, N. A., Joseph, C. L., Allen, G., Benson, A. K., and Mills, D. A. (2012). Next-generation sequencing reveals significant bacterial diversity of botrytized wine. *PLoS ONE* 7:e36357. doi: 10.1371/journal.pone.0036357
- Bokulich, N. A., and Mills, D. A. (2012). Differentiation of mixed lactic acid bacteria communities in beverage fermentations using targeted terminal restriction fragment length polymorphism. *Food Microbiol.* 31, 126–132. doi: 10.1016/j.frm.2012.02.007
- Bokulich, N. A., and Mills, D. A. (2013). Improved selection of internal transcribed spacer-specific primers enables quantitative, ultra-high-throughput profiling of fungal communities. *Appl. Environ. Microbiol.* 79, 2519–2526. doi: 10.1128/AEM.03870-12
- Bokulich, N. A., Thorngate, J. H., Richardson, P. M., and Mills, D. A. (2014). Microbial biogeography of wine grapes is conditioned by cultivar, vintage,
- and climate. *Proc. Natl. Acad. Sci. U.S.A.* 111, E139–E148. doi: 10.1073/pnas.1317377110
- Burns, K. N., Bokulich, N. A., Cantu, D., Greenhut, R. F., Kluepfel, D. A., O’Geen, A. T., et al. (2016). Vineyard soil bacterial diversity and composition revealed by 16S rRNA genes: differentiation by vineyard management. *Soil Biol. Biochem.* 103, 337–348. doi: 10.1016/j.soilbio.2016.09.007
- Burns, K. N., Kluepfel, D. A., Strauss, S. L., Bokulich, N. A., Cantu, D., and Steenwerth, K. L. (2015). Vineyard soil bacterial diversity and composition revealed by 16S rRNA genes: differentiation by geographic features. *Soil Biol. Biochem.* 91, 232–247. doi: 10.1016/j.soilbio.2015.09.002
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth. f.303
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624. doi: 10.1038/ismej.2012.8
- Compani, S., Clément, C., and Sessitsch, A. (2010). Plant growth-promoting bacteria in the rhizo-and endosphere of plants: their role, colonization, mechanisms involved and prospects for utilization. *Soil Biol. Biochem.* 42, 669–678. doi: 10.1016/j.soilbio.2009.11.024
- Courtois, S., Frostegård, Å., Göransson, P., Depret, G., Jeannin, P., and Simonet, P. (2001). Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environ. Microbiol.* 3, 431–439. doi: 10.1046/j.1462-2920.2001.00208.x
- Curtis, T. P. (2002). Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. U.S.A.* 99, 10494–10499. doi: 10.1073/pnas.142680199
- David, V., Terrat, S., Herzine, K., Claisse, O., Rousseaux, S., Tourdot-Maréchal, R., et al. (2014). High-throughput sequencing of amplicons for monitoring yeast biodiversity in must and during alcoholic fermentation. *J. Ind. Microbiol. Biotechnol.* 41, 811–821. doi: 10.1007/s10295-014-1427-2
- Divol, B., and Lonvaud-Funel, A. (2005). Evidence for viable but nonculturable yeasts in botrytis affected wine. *J. Appl. Microbiol.* 99, 85–93. doi: 10.1111/j.1365-2672.2005.02578.x
- Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604
- Edgar, R. C. (2016). SINTAX, a Simple Non-Bayesian Taxonomy Classifier for 16S and ITS Sequences. Available at: <http://biorxiv.org/content/early/2016/09/09/074161> doi: 10.1101/074161
- Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/210X.12114
- Eren, A. M., Sogin, M. L., and Maignien, L. (2016). Editorial: new insights into microbial ecology through subtle nucleotide variation. *Front. Microbiol.* 7:1318. doi: 10.3389/fmicb.2016.01318
- Fontaine, F., Pinto, C., Vallet, J., Clément, C., Gomes, A. C., and Spagnolo, A. (2016). The effects of grapevine trunk diseases (GTDs) on vine physiology. *Eur. J. Plant Pathol.* 144, 707–721. doi: 10.1007/s10658-015-0770-0
- Gilbert, J. A., Meyer, F., Jansson, J., Gordon, J., Pace, N., Tiedje, J., et al. (2010). The earth microbiome project: meeting report of the “1 st EMP meeting on sample selection and acquisition” at argonne national laboratory October 6th 2010. *Stand. Genomic Sci.* 3, 249–253. doi: 10.4056/ags.1443528
- Grangeteau, C., Roullier-Gall, C., Rousseaux, S., Gougeon, R. D., Schmitt-Kopplin, P., Alexandre, H., et al. (2017). Wine microbiology is driven by

- vineyard and winery anthropogenic factors. *Microb. Biotechnol.* 10, 354–370. doi: 10.1111/1751-7915.12428
- Greuter, D., Loy, A., Horn, M., and Rattei, T. (2016). probeBase—an online resource for rRNA-targeted oligonucleotide probes and primers: new features. *Nucleic Acids Res.* 44, D586–D589. doi: 10.1093/nar/gkv1232
- Haggerty, J. M., and Dinsdale, E. A. (2017). Distinct biogeographical patterns of marine bacterial taxonomy and functional genes. *Glob. Ecol. Biogeogr.* 26, 177–190. doi: 10.1111/geb.12528
- Herrick, J. B., Madsen, E., Batt, C., and Ghiorse, W. (1993). Polymerase chain reaction amplification of naphthalene-catabolic and 16S rRNA gene sequences from indigenous sediment bacteria. *Appl. Environ. Microbiol.* 59, 687–694.
- Hilger, A., and Myrold, D. (1991). Method for extraction of Frankia DNA from soil. *Agric. Ecosyst. Environ.* 34, 107–113. doi: 10.1016/0167-8809(91)90098-I
- Holben, W. E. (1994). “Isolation and purification of bacterial DNA from soil,” in *Methods of Soil Analysis: Part 2—Microbiological and Biochemical Properties*, eds R. Weaver, P. Bottomly, and S. Angle (Madison, WI: Soil Science Society of America), 727–751.
- Hong, S., Bunge, J., Leslin, C., Jeon, S., and Epstein, S. S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* 3, 1365–1373. doi: 10.1038/ismej.2009.89
- Keisam, S., Romi, W., Ahmed, G., and Jeyaram, K. (2016). Quantifying the biases in metagenome mining for realistic assessment of microbial ecology of naturally fermented foods. *Sci. Rep.* 6:34155. doi: 10.1038/srep34155
- Kennedy, J. (2002). Understanding grape berry development. *Prac. Winery Vineyard* 24, 14–23.
- Khandelwal, R. A., Olivier, B. G., Röling, W. F., Teusink, B., and Bruggeman, F. J. (2013). Community flux balance analysis for microbial consortia at balanced growth. *PLoS ONE* 8:e64567. doi: 10.1371/journal.pone.0064567
- Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41, e1. doi: 10.1093/nar/gks808
- Knight, S., Klaere, S., Fedrizzi, B., and Goddard, M. R. (2015). Regional microbial signatures positively correlate with differential wine phenotypes: evidence for a microbial aspect to terroir. *Sci. Rep.* 5:14233. doi: 10.1038/srep14233
- Lee, C. K., Herbold, C. W., Polson, S. W., Wommack, K. E., Williamson, S. J., McDonald, I. R., et al. (2012). Groundtruthing next-gen sequencing for microbial ecology—biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS ONE* 7:e44224. doi: 10.1371/journal.pone.0044224
- Leveau, J., and Tech, J. (2010). “Grapevine microbiomics: bacterial diversity on grape leaves and berries revealed by high-throughput sequence analysis of 16S rRNA amplicons,” in *Proceedings of the International Symposium on Biological Control of Postharvest Diseases: Challenges and Opportunities*, Vol. 905, Leesburg, VA, 31–42.
- Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., et al. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* 16, 2659–2671. doi: 10.1111/1462-2920.12250
- Louca, S., and Doebeli, M. (2015). Calibration and analysis of genome-based models for microbial ecology. *eLife* 4:e08208. doi: 10.7554/eLife.08208
- López-Rituer, E., Savorani, F., Avenoza, A., Bustos, J. S. H., Peregrina, J. S. M., and Engelsen, S. B. (2012). Investigations of La Rioja terroir for wine production using 1H NMR metabolomics. *J. Agric. Food Chem.* 60, 3452–3461. doi: 10.1021/jf304361d
- Lugtenberg, B., and Kamlova, F. (2009). Plant-growth-promoting rhizobacteria. *Annu. Rev. Microbiol.* 63, 541–556. doi: 10.1146/annurev.micro.62.081307.162918
- Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D., and Dangl, J. L. (2013). Practical innovations for high-throughput amplicon sequencing. *Nat. Methods* 10, 999–1002. doi: 10.1038/nmeth.2634
- Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2:e593. doi: 10.7717/peerj.593
- Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 3:e1420. doi: 10.7717/peerj.1420
- Martins, G., Lauga, B., Miot-Sertier, C., Mercier, A., Lonvaud, A., Soulas, M.-L., et al. (2013). Characterization of epiphytic bacterial communities from grapes, leaves, bark and soil of grapevine plants grown, and their relations. *PLoS ONE* 8:e73013. doi: 10.1371/journal.pone.0073013
- McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139
- McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531
- Miller, D., Bryant, J., Madsen, E., and Ghiorse, W. (1999). Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. *Appl. Environ. Microbiol.* 65, 4715–4724.
- Millet, V., and Lonvaud-Funel, A. (2000). The viable but non-culturable state of microorganisms during storage. *Lett. Appl. Microbiol.* 30, 136–141. doi: 10.1046/j.1472-765x.2000.00684.x
- Neeley, E. T., Phister, T. G., and Mills, D. A. (2005). Differential real-time PCR assay for enumeration of lactic acid bacteria in wine. *Appl. Environ. Microbiol.* 71, 8954–8957. doi: 10.1128/AEM.71.12.8954-8957.2005
- Neilson, J. W., Jordan, F. L., and Maier, R. M. (2013). Analysis of artifacts suggests DGGE should not be used for quantitative diversity analysis. *J. Microbiol. Methods* 92, 256–263. doi: 10.1016/j.mimet.2012.12.021
- OIV (2015). *World Vitiviniculture Situation 2015*. Paris: Office international de la vigne et du vin.
- Parada, A. E., Needham, D. M., and Fuhrman, J. A. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* 18, 1403–1414. doi: 10.1111/1462-2920.13023
- Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658
- Petric, I., Philippot, L., Abbate, C., Bispo, A., Chesnot, T., Hallin, S., et al. (2011). Inter-laboratory evaluation of the ISO standard 11063 “Soil quality—Method to directly extract DNA from soil samples”. *J. Microbiol. Methods* 84, 454–460. doi: 10.1016/j.mimet.2011.01.016
- Piao, H., Hawley, E., Kopf, S., DeScenzo, R., Sealock, S., Henick-Kling, T., et al. (2015). Insights into the bacterial community and its temporal succession during the fermentation of wine grapes. *Front. Microbiol.* 6:809. doi: 10.3389/fmicb.2015.00809
- Picard, C., Ponsonnet, C., Paget, E., Nesme, X., and Simonet, P. (1992). Detection and enumeration of bacteria in soil by direct DNA extraction and polymerase chain reaction. *Appl. Environ. Microbiol.* 58, 2717–2722.
- Pinto, A. J., and Raskin, L. (2012). PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS ONE* 7:e43093. doi: 10.1371/journal.pone.0043093
- Pinto, C., Pinho, D., Cardoso, R., Custodio, V., Fernandes, J., Sousa, S., et al. (2015). Wine fermentation microbiome: a landscape from different Portuguese wine appellations. *Front. Microbiol.* 6:905. doi: 10.3389/fmicb.2015.00905
- Pinto, C., Pinho, D., Sousa, S., Pinheiro, M., Egas, C., and Gomes, A. C. (2014). Unravelling the diversity of grapevine microbiome. *PLoS ONE* 9:e85622. doi: 10.1371/journal.pone.0085622
- Porteous, L., Seidler, R., and Watrud, L. (1997). An improved method for purifying DNA from soil for polymerase chain reaction amplification and molecular ecology applications. *Mol. Ecol.* 6, 787–791. doi: 10.1046/j.1365-294X.1997.00241.x
- Portillo, M. D. C., and Mas, A. (2016). Analysis of microbial diversity and dynamics during wine fermentation of Grenache grape variety by high-throughput barcoding sequencing. *Food Sci. Technol. LEB* 72, 317–321. doi: 10.1016/j.lwt.2016.05.009
- Ranjard, L., Poly, F., Combrisson, J., Richaume, A., and Nazaret, S. (1998). A single procedure to recover DNA from the surface or inside aggregates and in various size fractions of soil suitable for PCR-based assays of bacterial communities. *Eur. J. Soil Biol.* 34, 89–97. doi: 10.1016/S1164-5563(99)90006-7
- Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4, e2584. doi: 10.7717/peerj.2584

- Salma, M., Rousseaux, S., Sequeira-Le Grand, A., Divol, B., and Alexandre, H. (2013). Characterization of the Viable but Nonculturable (VBNC) state in *Saccharomyces cerevisiae*. *PLoS ONE* 8:e77600. doi: 10.1371/journal.pone.0077600
- Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1186/s12915-014-0087-z
- Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* 43:e37. doi: 10.1093/nar/gku1341
- Schloss, P. D., Jenior, M. L., Koumpouras, C. C., Westcott, S. L., and Highlander, S. K. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ.* 4:e1869. doi: 10.7717/peerj.1869
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09
- Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60
- Seguritan, V., and Rohwer, F. (2001). FastGroup: a program to dereplicate libraries of 16S rDNA sequences. *BMC Bioinformatics* 2:9. doi: 10.1186/1471-2105-2-9
- Singer, E., Bushnell, B., Coleman-Derr, D., Bowman, B., Bowers, R. M., Levy, A., et al. (2016). High-resolution phylogenetic microbial community profiling. *ISME J.* 10, 2020–2032. doi: 10.1038/ismej.2015.249
- Sipos, R., Székely, A. J., Palatinszky, M., Révész, S., Márialigeti, K., and Nikolausz, M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.* 60, 341–350. doi: 10.1111/j.1574-6941.2007.00283.x
- Soergel, D. A., Dey, N., Knight, R., and Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J.* 6, 1440–1444. doi: 10.1038/ismej.2011.208
- Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., et al. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* 1:11. doi: 10.1186/2049-2618-1-11
- Stefanini, I., Albanese, D., Cavazza, A., Franciosi, E., De Filippo, C., Donati, C., et al. (2016). Dynamic changes in microbiota and mycobiota during spontaneous 'Vino Santo Trentino' fermentation. *Microb. Biotechnol.* 9, 195–208. doi: 10.1111/1751-7915.12337
- Tikhonov, M., Leach, R. W., and Wingreen, N. S. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J.* 9, 68–80. doi: 10.1038/ismej.2014.117
- Tofalo, R., Scirone, M., Corsetti, A., and Suzzi, G. (2012). Detection of *Brettanomyces* spp. in red wines using Real-Time PCR. *J. Food Sci.* 77, 545–549. doi: 10.1111/j.1750-3841.2012.02871.x
- Urakawa, H., Martens-Habbema, W., and Stahl, D. A. (2010). High abundance of ammonia-oxidizing Archaea in coastal waters, determined using a modified DNA extraction method. *Appl. Environ. Microbiol.* 76, 2129–2135. doi: 10.1128/AEM.02692-09
- Vandenkoornhuysse, P., Quaiser, A., Duhamel, M., Le Van, A., and Dufresne, A. (2015). The importance of the microbiome of the plant holobiont. *New Phytol.* 206, 1196–1206. doi: 10.1111/nph.13312
- Varma, A., and Palsson, B. O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* 60, 3724–3731.
- Verginer, M., Leitner, E., and Berg, G. (2010). Production of volatile metabolites by grape-associated microorganisms. *J. Agric. Food Chem.* 58, 8344–8350. doi: 10.1021/jf100393w
- Wang, C., García-Fernández, D., Mas, A., and Esteve-Zarzoso, B. (2015). Fungal diversity in grape must and wine fermentation assessed by massive sequencing, quantitative PCR and DGGE. *Front. Microbiol.* 6:1156. doi: 10.3389/fmicb.2015.01156
- Wagner, M. R., Lundberg, D. S., Coleman-Derr, D., Tringe, S. G., Dangl, J. L., and Mitchell-Olds, T. (2014). Natural soil microbes alter flowering phenology and the intensity of selection on flowering time in a wild *Arabidopsis* relative. *Ecol. Lett.* 17, 717–726. doi: 10.1111/ele.12276
- Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681. doi: 10.1038/ismej.2015.235
- Yamamoto, H. (2000). Viable but nonculturable state as a general phenomenon of non-spore-forming bacteria, and its modeling. *J. Infect. Chemother.* 6, 112–114. doi: 10.1007/PL00012149
- Zarraonaindia, I., Owens, S. M., Weisenhorn, P., West, K., Hampton-Marcell, J., Lax, S., et al. (2015). The soil microbiome influences grapevine-associated microbiota. *MBio* 6, e02527-14. doi: 10.1128/mBio.02527-14
- Zomorrodi, A. R., and Maranas, C. D. (2012). OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol.* 8:e1002363. doi: 10.1371/journal.pcbi.1002363

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Belda, Zarraonaindia, Perisin, Palacios and Acedo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Perspective Study of Koumiss Microbiome by Metagenomics Analysis Based on Single-Cell Amplification Technique

Guoqiang Yao[†], Jie Yu[†], Qiangchuan Hou, Wenyan Hui, Wenjun Liu, Lai-Yu Kwok, Bilige Menghe, Tiansong Sun, Heping Zhang and Wenyi Zhang*

Key Laboratory of Dairy Biotechnology and Engineering, Ministry of Education, Inner Mongolia Agricultural University, Hohhot, China

OPEN ACCESS

Edited by:

Sabah Bidawid,
Health Canada, Canada

Reviewed by:

Kiliyuka Matthews Cilira,
Mount Kenya University, Kenya
Luca Cocolin,
University of Turin, Italy

*Correspondence:

Wenyi Zhang
zhangwenyizi@163.com

[†]These authors have contributed
equally to this work.

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 17 November 2016

Accepted: 23 January 2017

Published: 07 February 2017

Citation:

Yao G, Yu J, Hou Q, Hui W, Liu W, Kwok L-Y, Menghe B, Sun T, Zhang H and Zhang W (2017) A Perspective Study of Koumiss Microbiome by Metagenomics Analysis Based on Single-Cell Amplification Technique. *Front. Microbiol.* 8:165. doi: 10.3389/fmicb.2017.00165

Koumiss is a traditional fermented dairy product and a good source for isolating novel bacteria with biotechnology potential. In the present study, we applied the single-cell amplification technique in the metagenomics analysis of koumiss. This approach aimed at detecting the low-abundant bacteria in the koumiss. Briefly, each sample was first serially diluted until reaching the level of approximately 100 cells. Then, three diluted bacterial suspensions were randomly picked for further study. By analyzing 30 diluted koumiss suspensions, a total of 24 bacterial species were identified. In addition to the previously reported koumiss-associated species, such as *Lactobacillus (L.) helveticus*, *Lactococcus lactis*, *L. buchneri*, *L. kefiranciens*, and *Acetobacter pasteurianus*, we successfully detected three low-abundant taxa in the samples, namely *L. otakiensis*, *Streptococcus macedonicus*, and *Ruminococcus torques*. The functional koumiss metagenomes carried putative genes that relate to lactose metabolism and synthesis of typical flavor compounds. Our study would encourage the use of modern metagenomics to discover novel species of bacteria that could be useful in food industries.

Keywords: koumiss, metagenomics, single-cell amplification, bacterial diversity, low-abundant taxa

INTRODUCTION

Koumiss, also named chige, chigo, arrag, or airag, in the Mongolian language, is a type of traditional fermented dairy product. It has been a popular food in Mongolia and Inner Mongolia of China for centuries (Zhang and Zhang, 2011). People in these regions used to consume koumiss during grand festivities and sacrificial offerings (Zhang and Zhang, 2011). The earliest record of koumiss production can be traced back to the Han Dynasty (BC202–AD202). This product had attained widespread popularity during the Yuan Dynasty (AD1271–AD1368) (Zhang et al., 2010b). Nowadays, koumiss is a common food for the local people of Mongolia and Inner Mongolia, although only in few of these areas, it is manufactured in an industrial scale. Koumiss does not only provide rich nutrients, including high contents of essential amino acids and vitamins, but is also believed to relieve a wide range of medical conditions and is beneficial for postoperative care (Jagielski, 1877; Thompson, 1879).

Traditionally, koumiss is commonly produced in wooden casks, containers made of animal skin or urns. Fermentation occurs naturally at ambient temperature after the addition of filtrated

mare milk into the container with old koumiss, which serves as the starter culture (Zhang and Zhang, 2011). Koumiss is a good source of novel bacteria of biotechnology potential (Zhang et al., 2010a; Pan et al., 2011). Therefore, it is of intense interest to explore and preserve as many fermentation-associated koumiss bacteria as possible. During the last decades, a number of studies were performed to investigate the koumiss bacterial community (Wu et al., 2009; Hao et al., 2010), mainly studied by culture-, molecular biology- and pyrosequencing-based methods (Sun et al., 2010). Among these different approaches, the pyrosequencing-based method has provided the most comprehensive microbiota profile of koumiss independent from phenotypic traits and problems of cultivability of the individual microbes. However, the spectrum of functional genes coded by the koumiss bacteria and their fermentative capacities remain poorly characterized, particularly for the rare microbial populations.

The present study used the single cell genomics technique to analyze the bacterial metagenomes of 10 koumiss samples collected from Mongolia and Inner Mongolia of China. The current work has applied state-of-art technologies in investigating the bacterial diversity in dairy products. Our work has demonstrated the feasibility of discovering low-abundant taxa by applying the single cell metagenomics approach. The encouraging results would promote the development and application of novel approaches in tackling problems in a traditional field of research.

MATERIALS AND METHODS

Preparation of Samples

A total of 10 koumiss samples were collected from Mongolia (MG14, MG15, MG16, MG17, and MG18) and Inner Mongolia of China (NM17, NM18, NM19, NM20, and NM21) for the metagenomics study. Samples were collected aseptically and were transported in dry ice.

One milliliter of each sample was pretreated according to the methodology described in Ward et al. (2013) with some modifications. Briefly, the samples were thawed in an ice bath for 3–5 min. After the samples melted, they were subject to low speed centrifugation to remove impurities and eukaryotic cell clumps. Prokaryotic cells were then pelleted from the milk sera by centrifugation at $13,000 \times g$ for 15 min. The pellets were re-suspended in 2 mL phosphate buffered saline (PBS) with 1% Triton X-100 and incubated for 2 h at 37°C to lyse any remaining eukaryotic cells. Subsequently, bacteria were pelleted by centrifugation at $13,000 \times g$ for 15 min and the pellets were re-suspended in 500 μL PBS. Finally, the centrifugation step was repeated once more to wash the bacterial cells.

Gradient Dilution and Multiple Displacement Amplification

To detect the low-abundant bacteria, the bacterial suspension derived from each koumiss sample was serially diluted for subsequent amplification reaction. The cell number in each sample was roughly estimated under a microscope (Nikon,

Tokyo, Japan) using a cell counting chamber (Qiujing, Shanghai, China). The dilution step was continued until the cell number in each bacterial suspension reached approximately 100. Multiple displacement amplification of the diluted cells was performed using the REPLI-g Single Cell Kit (Qiagen, Germantown, MD, USA) according to the manufacturer's instructions.

Library Construction and Sequencing

Amplified DNA was sheared randomly, and the fragments of approximately 500 bp were selected. After the library construction, PerkinElmer LabChip[®] GX Touch and StepOnePlusTM Real-Time PCR System were used for library quality inspection. Finally, 125-bp paired-end reads were sequenced on the Illumina HiSeq 2500 platform according to the manufacturer's instructions.

Data Analysis

Sequence Quality Check and Filtering

Raw reads generated by the sequencer might contain artificial reads of adapter contamination during the library construction. Therefore, three steps were performed to obtain a high-quality clean read dataset: (1) elimination of reads caused by adapter contamination; (2) removal of reads with an average score below a phred score of Q30, which was considered as the lowest cutoff for a high-quality base; (3) removal of reads with a significant excess of "N" ($\geq 5\%$ of the read). The downstream analysis was based on the clean data. Moreover, the statistical base quality, based on Q30 and the GC content, were calculated.

Alignments against the host genome were carried out to remove the host-originated contaminant sequences. Any host-originated reads were discarded before further comparison with bacteria (or viruses) genome reference sequences. To obtain more accurate results, the Burrows–Wheeler aligner (BWA) (Version 0.97a) MEM model was used in the alignments (Li and Durbin, 2009).

Taxonomic Assignment and Diversity

The web software Metaphlan was used for taxonomic assignment to genus and species levels (Segata et al., 2012). To compare the diversity of species within and between samples, we analyzed the alpha- and beta-diversity by the R-related package.

Read Assembly, Gene Prediction, and Annotation

To obtain more comprehensive information, we assembled the sheared fragments into genome (contigs). However, due to the presence of multiple species, which is common in metagenomic samples, we improved the bioinformatics genome assembling method normally used for single species analysis by integrating SPAdes (Version 3.6.2), in-house scripts, and metagenomic databases (Zerbino and Birney, 2008; Nurk et al., 2013).

The MetaGeneMark software was used for gene prediction of the assembled contigs (Noguchi et al., 2006). Redundant genes were removed using CD-HIT with the coverage of 90 and 95% identity (Li and Godzik, 2006; Fu et al., 2012). Relative abundances of the genes were determined by aligning high-quality sequencing reads to the gene catalog using the same procedure. The downstream discrepant analysis was based on

the gene relative abundances. Gene annotation was performed by aligning the high-quality sequences against several public databases (namely NCBI non-redundant database, NR; Clusters of Orthologous Groups of proteins, COGs; Kyoto Encyclopedia of Genes and Genomes, KEGG) using BLAST (Altschul et al., 1997). A domain search was performed by using Interproscan (Mulder and Apweiler, 2008).

Nucleotide Sequence Accession Numbers

The sequence data reported in this study have been deposited in the SRA database (Accession No.SRP083102).

RESULTS

Experimental Design and Sequencing

To detect the low-abundant bacteria in koumiss, the single-cell amplification technique was used to analyze the metagenomes of the samples. Three bacterial suspensions, each of approximately 100 cells, were derived from an independent koumiss sample by serial dilution. A total of 30 diluted suspensions were analyzed. Each diluted sample was given a different sample code, i.e., the sample identity number followed by 1, 2, or 3, representing the three separate dilutions. Based on a premise that some rare species will be present in one of the dilutions, multiple displacement amplification of the cells was carried out; and around 5 Gb data were generated for every koumiss bacterial suspension.

A total of 1,040,323,864 raw reads were generated from the 10 koumiss samples (a total of 30 bacterial suspensions). The average number of the reads for each 100-cell-suspension was 34,677,462 (Supplementary Table S1). After trimming and filtering of the unqualified sequences, we obtained 1,018,381,702 clean reads for all samples (Supplementary Table S1). The values of Shannon index, Simpson index, Chao1 index, and the number of observed species (Figures 1–4) showed that most koumiss samples had a high bacterial biodiversity. The Shannon–Wiener diversity curves showed that the sequence depth was adequate for all samples (Figure 1).

Taxonomic Annotation

The high-quality sequences were assigned to different taxonomic levels to enable an in-depth analysis of the sample bacterial communities. With reference to some published studies on koumiss biodiversity (Wu et al., 2009; Hao et al., 2010; Sun et al., 2010), we classified the known and previously not reported koumiss-associated bacteria as common and rare taxa, respectively.

The high-quality sequences represented 13 different genera (Figure 5). Three of them had an average relative abundance of over 1%, including *Lactobacillus* (L.), *Lactococcus*, and *Streptococcus*. Particularly, *Lactobacillus* and *Lactococcus* were the two most abundant genera found in the koumiss. The proportion of *Lactobacillus* in the samples ranged from 52.72 to 99.96%. Two members of this genus, *L. helveticus* and *L. kefirnafaciens*, were dominated among most koumiss samples (Figure 6). The species *L. buchneri* was present mostly in the Mongolian samples, while

Lactococcus lactis was detected in most samples regardless of the sampling region.

Metagenomic Assembly, Gene Prediction, and Functional Annotation

Assembling of the reads resulted in an assembly length of 614,392,623 bp. The N50 values for the assemblies ranged from 5,596 to 35,200 bp (Supplementary Table S2). The number of predicted genes in the koumiss samples ranged from 10,347 to 34,547, with an average length ranging from 647.82 to 985.33 bp (Supplementary Table S3). Although empirical gene functional analyses were beyond the scope of the current study, we annotated the koumiss bacterial microbiomes using the COG and KEGG databases, which predicted gene function largely based on sequence homology.

A total of 545 matches in the annotation output showed high homology to the lactose utilization genes (COGs category of carbohydrate and metabolism, G), and some of them were located within the same contig (Table 1).

Some other sequences might code for putative genes within the COGs category of amino acid transport and metabolism (E), including sequences that corresponded to casein-degrading proteinases, the Opp system for taking up oligopeptides of 4–18 residues, and aminopeptidases (e.g., leucyl aminopeptidase, peptidyl-dipeptidase A, aminopeptidase N, proline iminopeptidase, and endopeptidase). Although some sequences shared high similarities with the aminotransferases specific for arginine, aspartate, methionine, and isoleucine, only one significant match was identified, which corresponded to a putative *S. macedonicus*-originated class I/class II domain (IPR004839)-containing aminotransferase specific for tyrosine and phenylalanine. Finally, a number of sequences shared high homology to the amino acid lyases, including S-ribosylhomocysteine lyase, argininosuccinate lyase, aspartate ammonia-lyase, cystathione gamma-lyase, histidine ammonia-lyase, and O-acetylhomoserine (thiol)-lyase.

DISCUSSION

Koumiss is a popular traditional fermented dairy product in Mongolia and Inner Mongolia of China. Although a number of studies have been performed to investigate the bacterial diversity in koumiss, little information has been obtained regarding the genetic capacity of the koumiss microbiota. Here, we applied the single-cell amplification technique to analyze the koumiss bacterial metagenomes, particularly focusing on the low-abundant bacterial population.

The typical metagenomics approach has previously been applied to describe the koumiss microbiota. However, due to the high cost of achieving a deep sequence, the rare microbiota population in the samples is often covered inadequately. Thus, both the phylogenetic and functional metagenomes of the minority bacteria remain limited. Our single-cell amplification method involved the serial dilution of samples to 100-cell suspensions. Owing to the low number of cells present in the diluted koumiss samples, only a limited amount of DNA would be

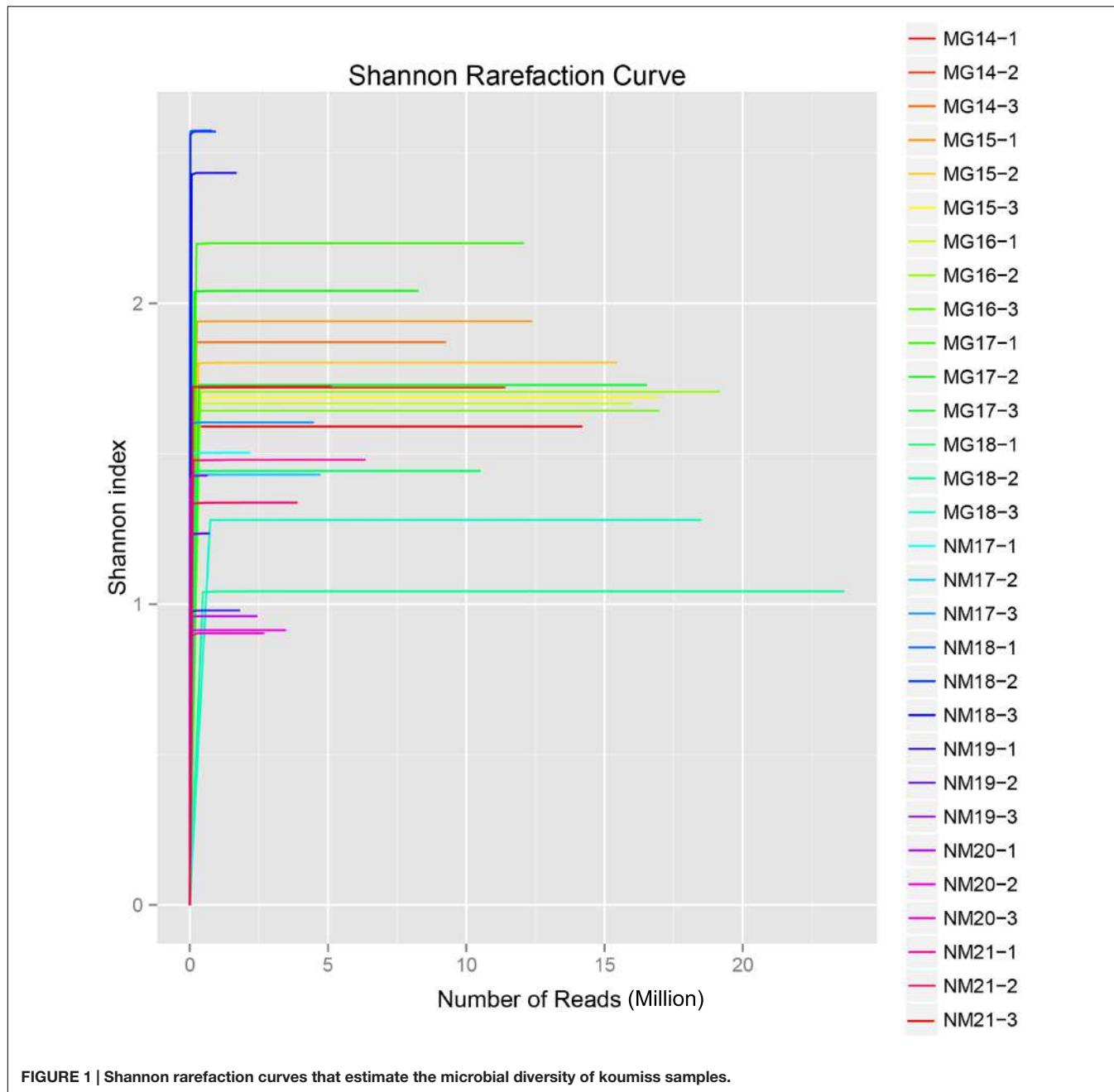


FIGURE 1 | Shannon rarefaction curves that estimate the microbial diversity of koumiss samples.

extracted. The amplification step increased the quantity of DNA materials to be analyzed and hence facilitated the metagenome analysis of samples containing minute DNA amounts. One drawback of this method was the difficulty in ensuring that sequences of each taxon would be equally amplified in the process; therefore, the results obtained here could only reflect the relative abundances of sequences but not absolute quantities of identified taxa or functional genes. Yet, this would have little effect on our analysis as the present study differs from other published works in focusing the rare bacterial population. Data generated by this work provide complementary information to the underrepresented population. We believe the present

approach is suitable for future analysis of bacteria diversity for different types of ecological environments.

The koumiss bacterial microbiota is mainly consisted of lactic acid bacteria (LAB) and some acetic acid bacteria (Zhang and Zhang, 2011). As expected, our dataset contained mostly sequences representing the LAB and acetic acid bacteria. Sequences corresponding to *Lactobacillus helveticus* were dominating across all samples. Besides, sequences representing the species, *L. kefirinofaciens*, *L. buchneri*, *L. kefirinofaciens*, *Enterococcus (E.) casseliflavus*, *E. faecalis*, *E. faecium*, *Leuconostoc mesenteroides*, *Lactococcus lactis*, and *Acetobacter pasteurianus*, were also found. The identification of sequences of different

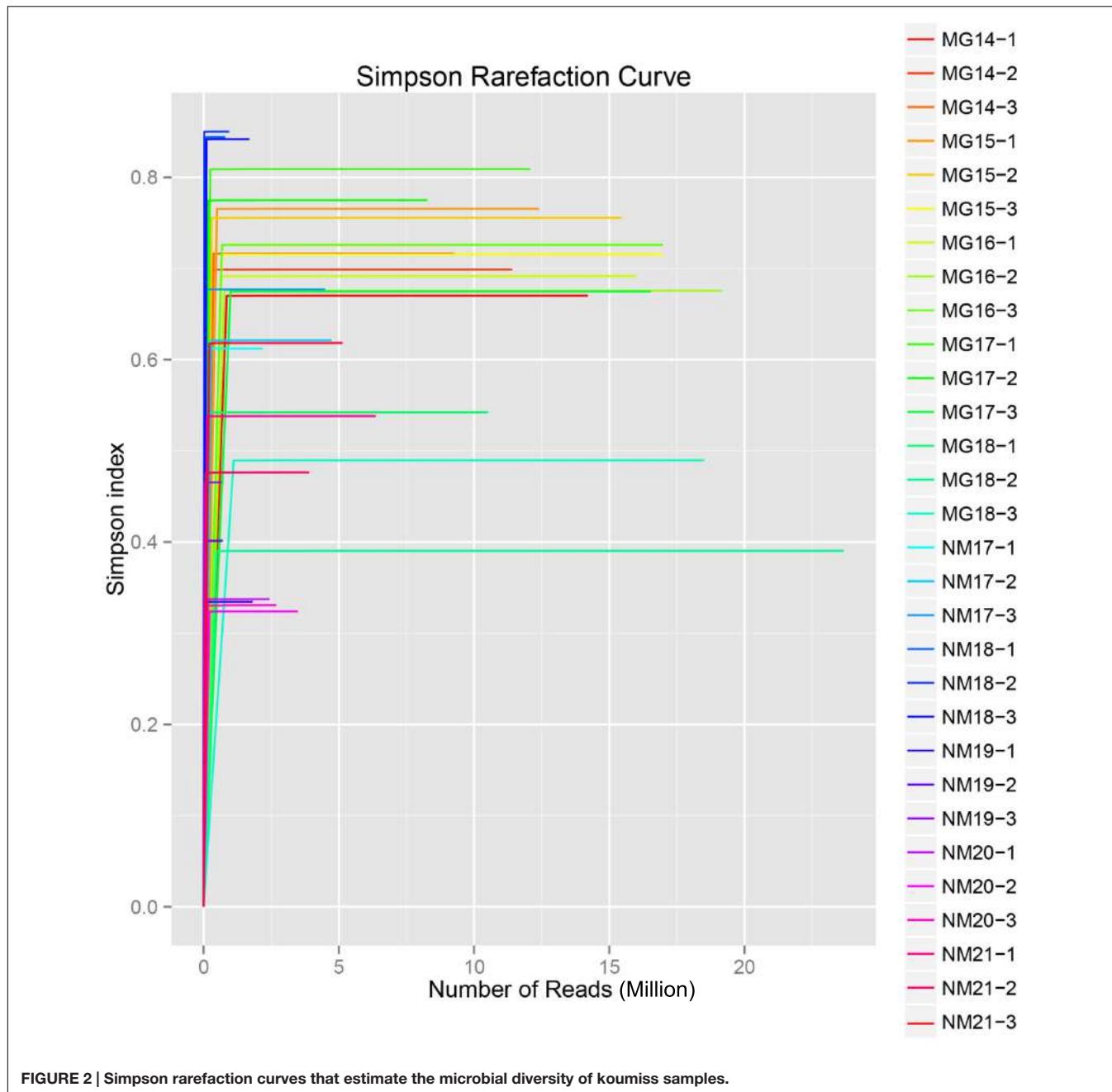


FIGURE 2 | Simpson rarefaction curves that estimate the microbial diversity of koumiss samples.

taxa may not be enough to show the viability of the bacteria, it nevertheless reflects the bacterial community at some point of the fermentation process. Miyamoto et al. (2010) suggests that the bacterial prevalence in the final fermented products is related to their acid stress tolerance. Generally, lactobacilli have a higher acid tolerance than lactococci, which may explain the high relative abundance of lactobacilli sequences present in our dataset. On the other hand, the frequent occurrence of *L. helveticus* in koumiss coincided with the observation of the dominance of *L. helveticus* sequences.

In our dataset, some sequences corresponded to *L. otakiensis*, which is a rare species that has never been reported in koumiss or

other dairy niches. The species was firstly described and isolated from the non-salted pickling solution used in producing sunki, a traditional Japanese pickle (Watanabe et al., 2009). It was discovered by amplified fragment length polymorphism profiling based on the *recA* gene. Since then, this species has not been reported to be associated with other food-related niches. Thus, it is likely that this bacterium belongs to the autochthonous flora of pickles. However, we cannot exclude the possibility that it was not found simply due to the sensitivity of detection method employed. *Lactobacillus otakiensis* can produce d-branched-chain amino acids, such as d-leucine, d-allo-isoleucine and d-valine (Doi et al., 2013). It has potential to be used in improving

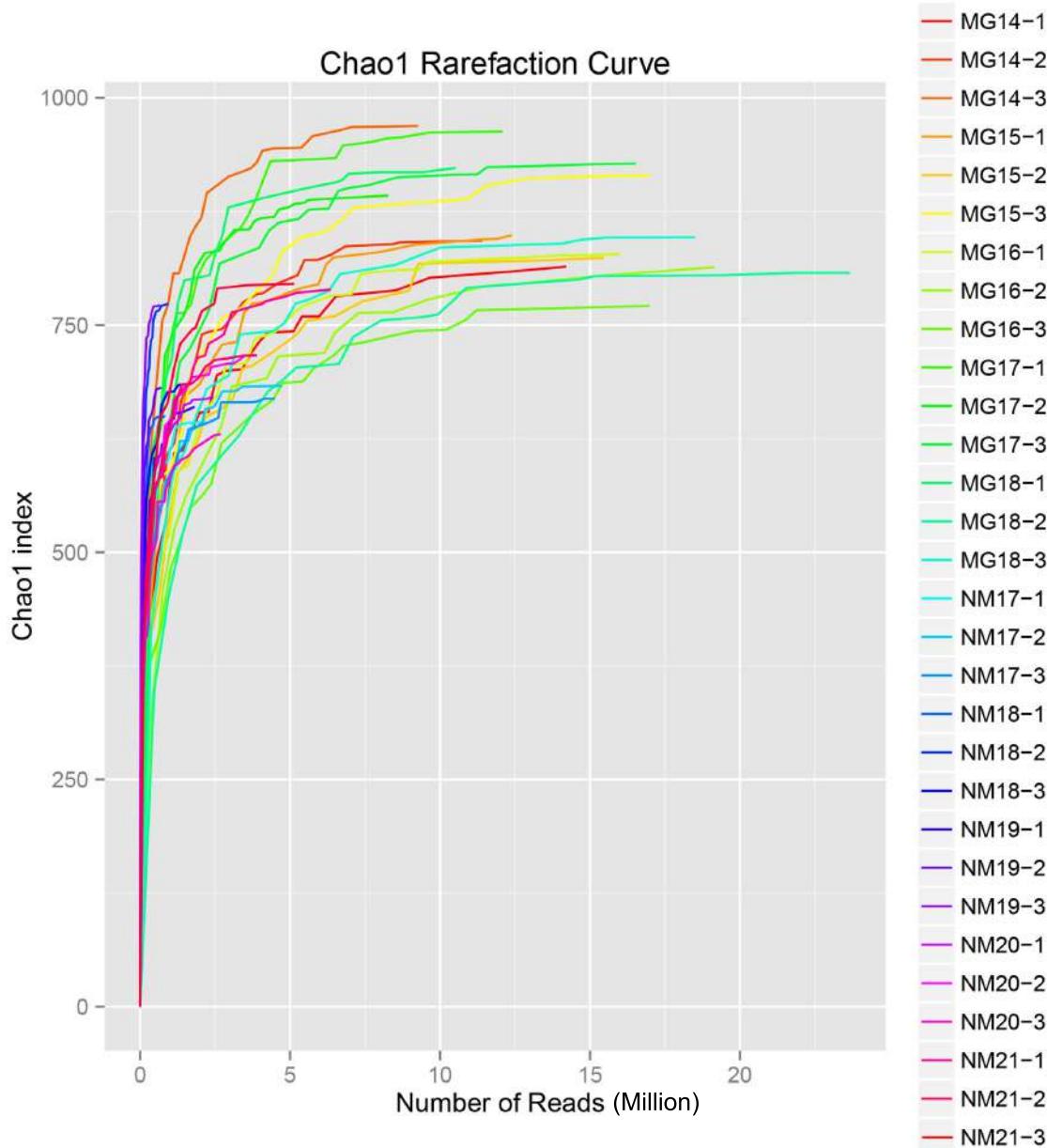


FIGURE 3 | Chao1 rarefaction curves that estimate the microbial diversity of koumiss samples.

production characteristics of certain fermented foods (Kato et al., 2011).

Our study also identified sequences representing the species *S. macedonicus*, which has never been reported as a koumiss-associated bacterium. Instead, it is a starter culture present in Greek sheep and goat cheeses (Georgalaki et al., 2000). Some members of this species are able to produce exopolysaccharides, bacteriocins (Vincent et al., 2001; Anastasiou et al., 2015), and gamma-aminobutyric acid (Franciosi et al., 2015). Even though this species is frequently isolated from fermented foods, the original niche of *S. macedonicus* has been controversial (Guarcello et al., 2016). Not until recently, Papadimitriou

et al. (2015) identified an acquired plasmid, pSMA198, from *Lactococcus lactis*. The plasmid was likely transferred via an ancestral genetic exchange event within a dairy product environment, hinting to the dairy origin of *S. macedonicus* (Papadimitriou et al., 2015). Similar to *S. thermophilus*, *S. macedonicus* is closely related to the commensals and opportunistic pathogens of the *S. bovis/S. equinus* complex.

The identification of sequences representing the species *Ruminococcus torques* was unexpected, as this bacterium is usually found in the gut environment. It is a normal human gut microbe that can degrade mucin oligosaccharides by constitutive production of secretory glycosidases (Hoskins et al., 1985).

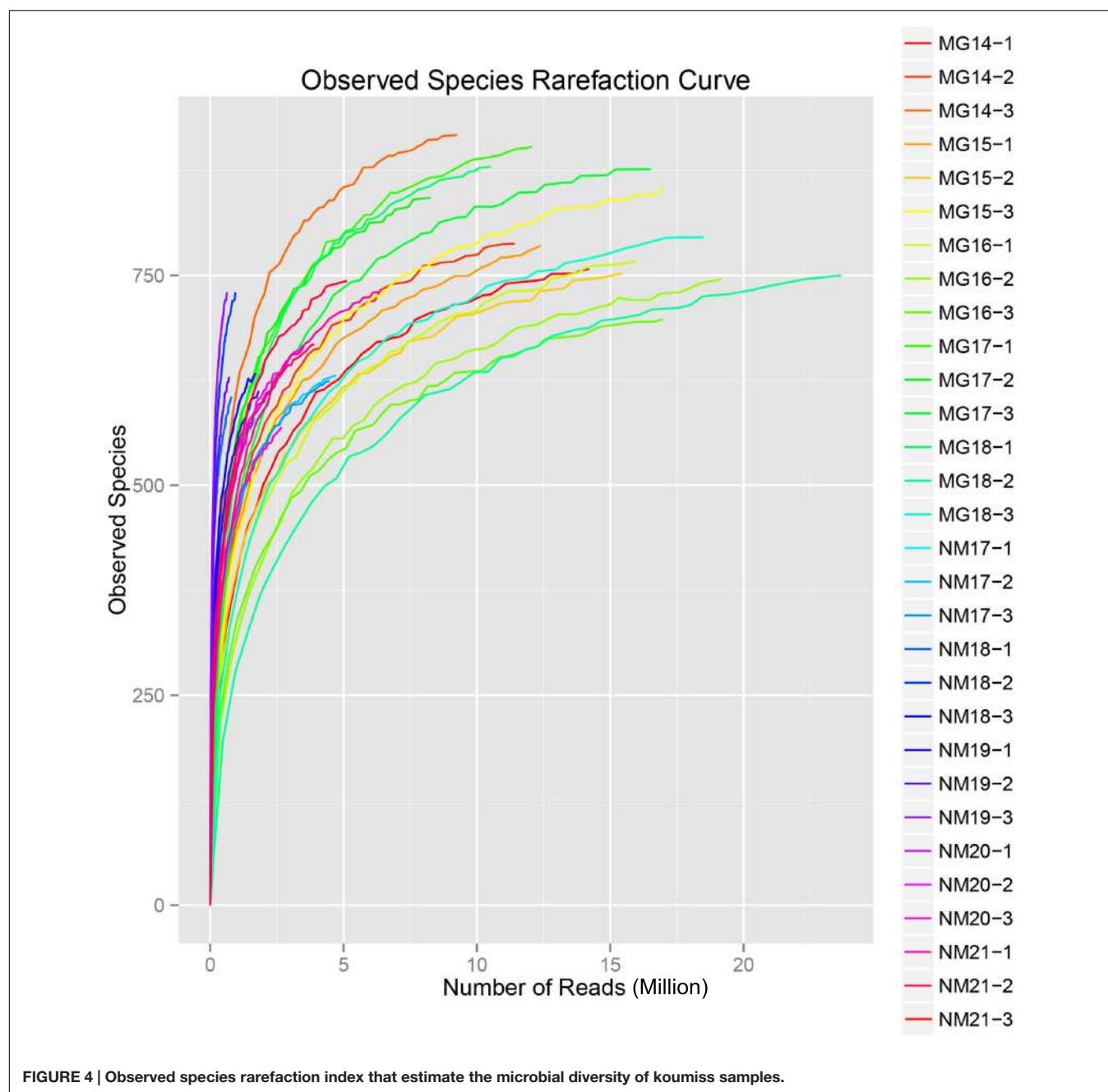


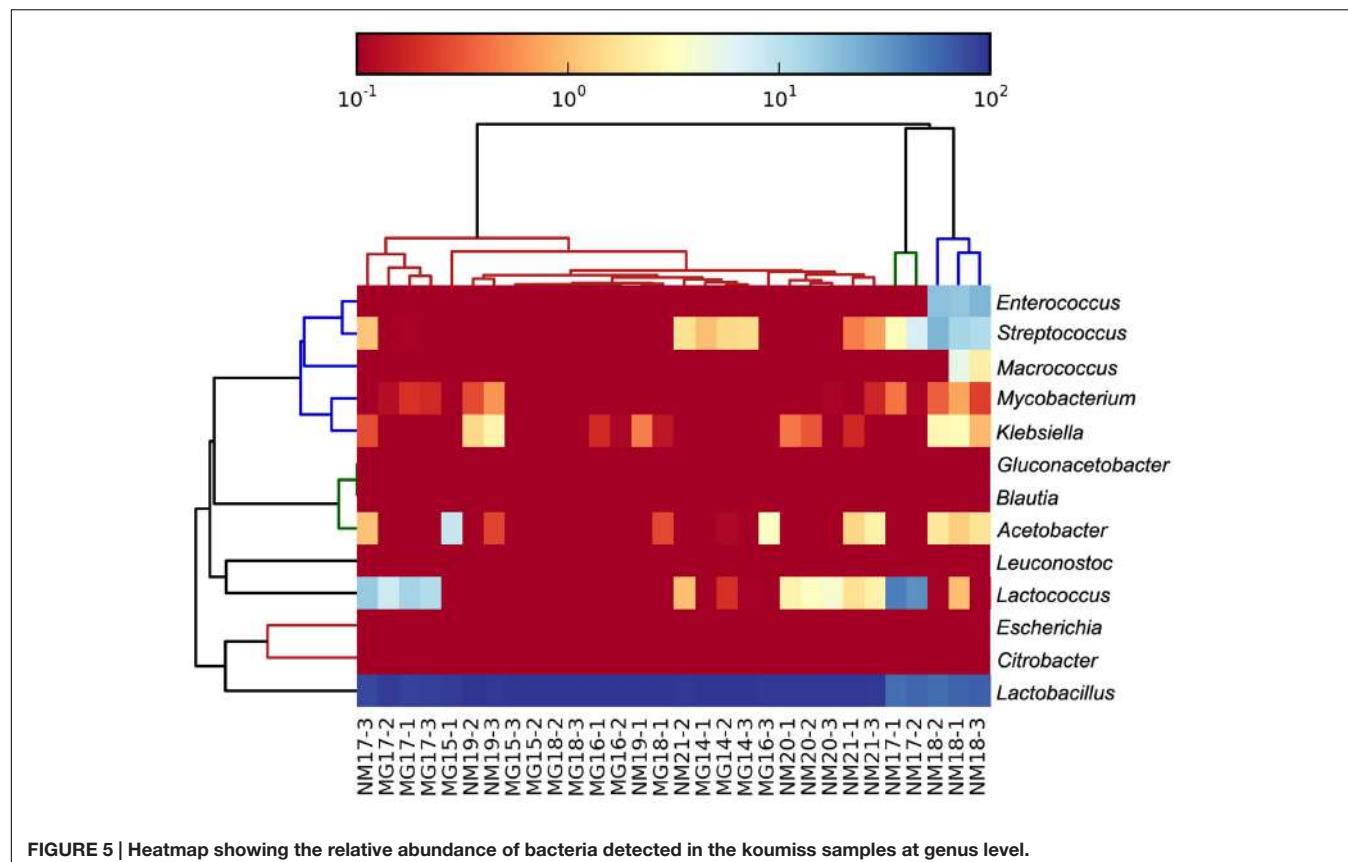
FIGURE 4 | Observed species rarefaction index that estimate the microbial diversity of koumiss samples.

Recent clinical evidence shows that the fecal abundance of this species is altered in children with autism spectrum disorder; however, its role in the disorder remains unclear (Wang et al., 2013).

Surprisingly, some of the sequences represented potential pathogens. For example, *Klebsiella pneumonia* is an opportunistic human pathogen that resides in around 40% of human and animal guts. *Mycobacterium orygis*, previously called the oryx bacillus, is a member of the *Mycobacterium tuberculosis* complex that may cause human tuberculosis (Dawson et al., 2012). These two species have been reported in raw milk but not koumiss. Therefore, their presence could ascribe to the contamination

during conventional koumiss production, especially under non- or low-aseptic manipulation conditions.

Bacterial metabolism plays an important role in forming the koumiss characteristics and quality. Microbial-based processes such as lipolysis and proteolysis are required for synthesizing koumiss aromatic and flavor compounds (Gesudu et al., 2016). Consistently, the bacterial metagenome contained sequences that potentially code for lactose degradation and proteolytic systems. Unlike the relatively simple lactose catabolic pathways, the LAB proteolytic systems are made up of a diverse array of enzymes (Chen et al., 2014). Compared with other detected koumiss LAB, the dominant species *L. helveticus* is characterized with a



high proteolytic activity (Zhang et al., 2015). Most LAB possess only one cell-envelope proteinase that initiates the milk casein hydrolysis, whereas *L. helveticus* contains at least two of these enzymes, namely *PrtH* and *PrtH2* (Zhao et al., 2011). Thus, the high proportions of sequences corresponding to the strong proteolytic species of *L. helveticus* and proteolysis-related genes may link to the relatively high contents of peptides and free amino acids in koumiss.

One difficulty in industrial scale koumiss production is the control of flavor perception, as koumiss has been traditionally made by natural fermentation. Thus, it has been hard to define the flavor and the profile of key flavor components, particularly in the presence of the natural contaminants. The production of key flavor components is a result of fermentative and enzymatic degradation of amino acids, such as the branched-chain amino acids, methionine and aromatic amino acids (Ardo, 2006). Examples of flavor components include aldehydes, organic acids and esters, which are formed by transaminase (AT)-pathway (Helinck et al., 2004). This pathway is initiated by a transaminase that catalyzes the conversion of an amino acid to its corresponding α -keto acid (Helinck et al., 2004). Some sequences in our dataset corresponded to putative branched-chain and aromatic amino acid aminotransferases. Particularly, we found a putative aromatic amino acid aminotransferase I in the contig of the koumiss-associated species, *S. macedonicus*. Since our current work only annotated the microbiome *in silico*, our data are not enough to show that these identified gene sequences were

indeed functional to produce the aforementioned koumiss flavor compounds. The presence of these genes nevertheless suggest that they are some possible candidates for such fermentative activities; yet, further experimental work would be required to elucidate their exact functional roles.

Moreover, we found sequences corresponding to other amino acid conversion pathways including amino acid lyases and threonine aldolase. The former enzyme catalyzes the conversion of methionine to methanethiol (Irmler et al., 2008), thus resulting in dimethyl disulfide and dimethyl trisulfide formation (Fernandez et al., 2000), while the latter one catalyzes the conversion of threonine to glycine and acetaldehyde (Ott et al., 2000). Similarly, merely locating these sequences within the koumiss microbiome cannot serve as definite evidence in their actual biological roles; future experimental confirmation will be required.

CONCLUSION

The present study used a modified metagenomics method to analyze the bacterial microbiome of koumiss samples collected from Mongolia and Inner Mongolia of China. We characterized both the phylogenetic and functional metagenomes of the rare species in koumiss; and our dataset reflects traits that are of biotechnology interest and potential. Our study has demonstrated for the first time the feasibility of incorporating

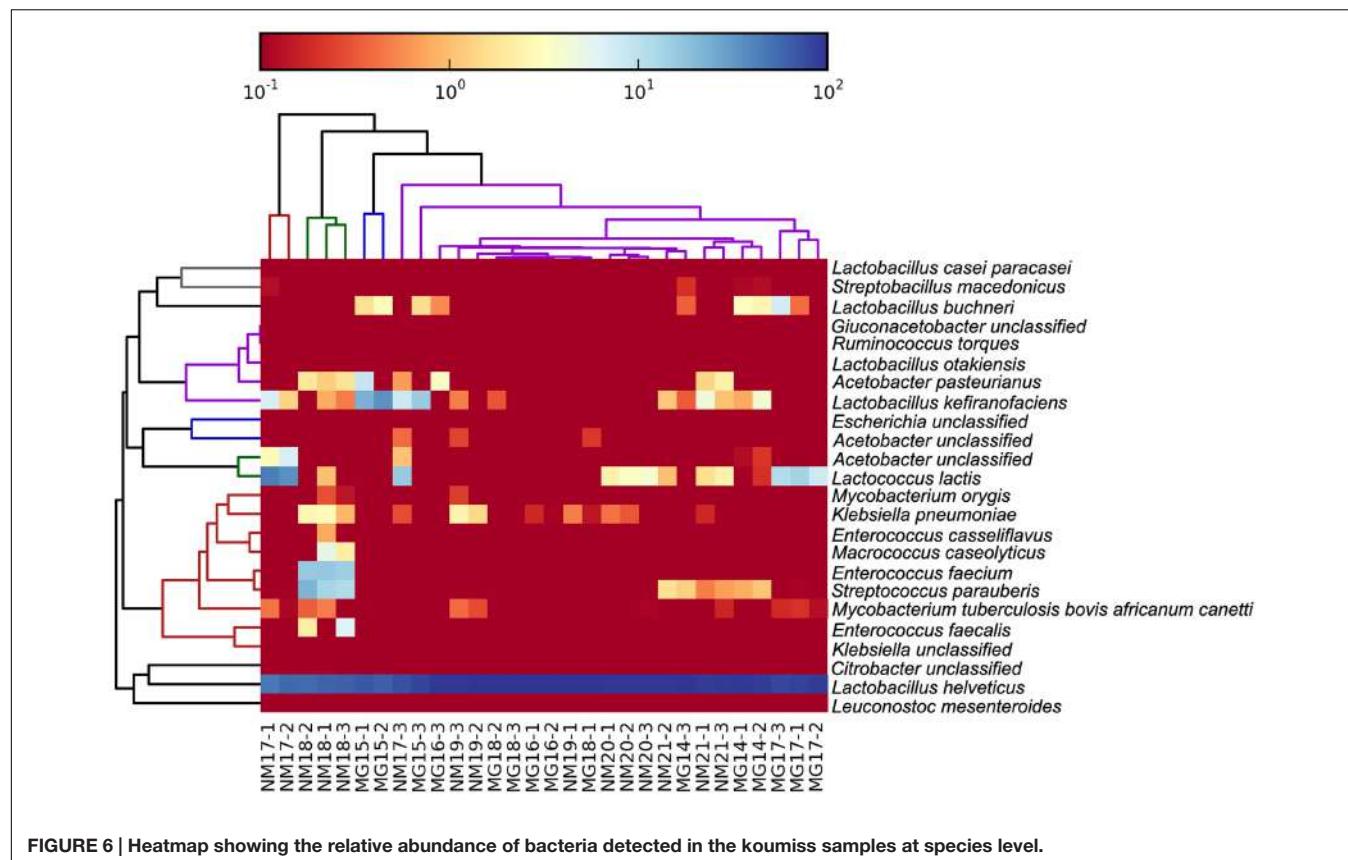


TABLE 1 | Lactose metabolism-related genes annotated in the koumiss bacterial metagenome.

Description	Gene name	EC numbers	Number of matches in the annotation output
Galactose-6-phosphate isomerase	lacA, lacB	EC:5.3.1.26	106
Tagatose 6-phosphate kinase	lacC	EC:2.7.1.144	12
Tagatose 1,6-diphosphate aldolase	lacD	EC:4.1.2.40	93
PTS system, lactose-specific IIA component	lacF (lacE)	EC:2.7.1.69	97
6-phospho-beta-galactosidase	lacG	E3.2.1.85	41
Beta-galactosidase	lacZ	EC:3.2.1.23	196

the single-cell amplification techniques in detecting koumiss bacterial microbiota, as well as microbial contaminants. The techniques developed herein can be used in future studies to monitor changes in the koumiss microbiome along the fermentation process, with focus on the minority microbial population. Furthermore, other omics techniques, such as transcriptomics, metabolomics, can be coupled to the current metagenomics analysis in order to confirm the functions and metabolic capacity of the koumiss microbiome.

Technically, the next step of this work would be to optimize the current method. For example, by increasing the sample dilution before DNA amplification, the chance of uncovering rare and novel bacteria may be improved. On the other hand, an alternative sequencing technique that can generate long reads, such as Pacific Biosciences single molecule, real-time sequencing technology, can be employed to improve the genome assembling process.

AUTHOR CONTRIBUTIONS

WZ and HZ designed the study. WZ, GY, JY, and L-YK wrote the manuscript. QH, WH, WL, BM, and TS performed experiments. WZ and QH analyzed data. All authors reviewed the manuscript.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (Grant No. 31571815).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.00165/full#supplementary-material>

REFERENCES

- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389
- Anastasiou, R., Driessche, G. V., Boutou, E., Kazou, M., Alexandraki, V., Vorgias, C. E., et al. (2015). Engineered strains of *Streptococcus macedonicus* towards an osmotic stress resistant phenotype retain their ability to produce the bacteriocin macedocin under hyperosmotic conditions. *J. Biotechnol.* 212, 125–133. doi: 10.1016/j.biote.2015.08.018
- Ardo, Y. (2006). Flavour formation by amino acid catabolism. *Biotechnol. Adv.* 24, 238–242. doi: 10.1016/j.biotechadv.2005.11.005
- Chen, Y. F., Zhao, W. J., Wu, R. N., Sun, Z. H., Zhang, W. Y., Wang, J. C., et al. (2014). Proteome analysis of *Lactobacillus helveticus* H9 during growth in skim milk. *J. Dairy Sci.* 97, 7413–7425. doi: 10.3168/jds.2014-8520
- Dawson, K. L., Bell, A., Kawakami, R. P., Coley, K., Yates, G., and Collins, D. M. (2012). Transmission of *Mycobacterium orygis* (*M. tuberculosis* complex species) from a tuberculosis patient to a dairy cow in New Zealand. *J. Clin. Microbiol.* 50, 3136–3138. doi: 10.1128/JCM.01652-12
- Doi, K., Mori, K., Mutaguchi, Y., Tashiro, K., Fujino, Y., Ohmori, T., et al. (2013). Draft genome sequence of D-branched-chain amino acid producer *Lactobacillus otakiensis* JCM 15040T, isolated from a traditional Japanese pickle. *Genome Announc.* 1, e546–e513. doi: 10.1128/genomeA.00546-13
- Fernandez, M., van Doesburg, W., Rutten, G. A., Marugg, J. D., Alting, A. C., van Kransenburgh, R., et al. (2000). Molecular and functional analyses of the metC gene of *Lactococcus lactis*, encoding cystathione beta-lyase. *Appl. Environ. Microbiol.* 66, 42–48. doi: 10.1128/AEM.66.1.42-48.2000
- Franciosi, E., Carafa, I., Nardin, T., Schiavon, S., Poznanski, E., Cavazza, A., et al. (2015). Biodiversity and gamma-aminobutyric acid production by lactic acid bacteria isolated from traditional alpine raw cow's milk cheeses. *Biomed. Res. Int.* 2015, 625740. doi: 10.1155/2015/625740
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565
- Georgalaki, M. D., Sarantinopoulos, P., Ferreira, E. S., De Vuyst, L., Kalantzopoulos, G., and Tsakalidou, E. (2000). Biochemical properties of *Streptococcus macedonicus* strains isolated from Greek Kasseri cheese. *J. Appl. Microbiol.* 88, 817–825. doi: 10.1046/j.1365-2672.2000.01055.x
- Gesdu, Q. M., Zheng, Y., Xi, X. X., Hou, Q. C., Xie, H. Y., Huang, W. Q., et al. (2016). Investigating bacterial population structure and dynamics in traditional koumiss from Inner Mongolia using single molecule real-time sequencing. *J. Dairy Sci.* 99, 7852–7863. doi: 10.3168/jds.2016-11167
- Guarcello, R., Carpino, S., Gaglio, R., Pino, A., Rapisarda, T., Caggia, C., et al. (2016). A large factory-scale application of selected autochthonous lactic acid bacteria for PDO Pecorino Siciliano cheese production. *Food Microbiol.* 59, 66–75. doi: 10.1016/j.fm.2016.05.011
- Hao, Y., Zhao, L., Zhang, H., Zhai, Z., Huang, Y., Liu, X., et al. (2010). Identification of the bacterial biodiversity in koumiss by denaturing gradient gel electrophoresis and species-specific polymerase chain reaction. *J. Dairy Sci.* 93, 1926–1933. doi: 10.3168/jds.2009-2822
- Helinck, S., Le Bars, D., Moreau, D., and Yvon, M. (2004). Ability of thermophilic lactic acid bacteria to produce aroma compounds from amino acids. *Appl. Environ. Microbiol.* 70, 3855–3861. doi: 10.1128/AEM.70.7.3855-3861.2004
- Hoskins, L. C., Agustines, M., McKee, W. B., Boulding, E. T., Kriaris, M., and Niedermeyer, G. (1985). Mucin degradation in human colon ecosystems. Isolation and properties of fecal strains that degrade ABH blood group antigens and oligosaccharides from mucin glycoproteins. *J. Clin. Invest.* 75, 944–953. doi: 10.1172/JCI111795
- Irmler, S., Raboud, S., Beisert, B., Rauhut, D., and Berthoud, H. (2008). Cloning and characterization of two *Lactobacillus casei* genes encoding a cystathione lyase. *Appl. Environ. Microbiol.* 74, 99–106. doi: 10.1128/AEM.00745-07
- Jagielski, V. (1877). The value of koumiss in the treatment of nausea, vomiting, and inability to retain other food on the stomach. *Br. Med. J.* 2, 919–921. doi: 10.1136/bmj.2.887.919
- Kato, S., Ishihara, T., Hemmi, H., Kobayashi, H., and Yoshimura, T. (2011). Alterations in D-amino acid concentrations and microbial community structures during the fermentation of red and white wines. *J. Biosci. Bioeng.* 111, 104–108. doi: 10.1016/j.jbosc.2010.08.019
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
- Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158
- Miyamoto, M., Seto, T., Nakajima, H., Burenjargal, S., Gombojav, A., Demberei, S., et al. (2010). Denaturing gradient gel electrophoresis analysis of lactic acid bacteria and yeasts in traditional Mongolian fermented milk. *Food Sci. Technol. Res.* 16, 319–326. doi: 10.3136/fstr.16.319
- Mulder, N. J., and Apweiler, R. (2008). The InterPro database and tools for protein domain analysis. *Curr. Protoc. Bioinformatics* Chapter 2, Unit2.7. doi: 10.1002/0471250953.bi0207s21
- Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 34, 5623–5630. doi: 10.1093/nar/gkl723
- Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., et al. (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* 20, 714–737. doi: 10.1089/cmb.2013.0084
- Ott, A., Germond, J. E., and Chaintreau, A. (2000). Origin of acetaldehyde during milk fermentation using (13)C-labeled precursors. *J. Agric. Food Chem.* 48, 1512–1517. doi: 10.1021/jf9904867
- Pan, D. D., Zeng, X. Q., and Yan, Y. T. (2011). Characterisation of *Lactobacillus fermentum* SM-7 isolated from koumiss, a potential probiotic bacterium with cholesterol-lowering effects. *J. Sci. Food. Agric.* 91, 512–518. doi: 10.1002/jsfa.4214
- Papadimitriou, K., Anastasiou, R., Maistrou, E., Plakas, T., Papandreou, N. C., Hamodrakas, S. J., et al. (2015). Acquisition through horizontal gene transfer of plasmid pSMA198 by *Streptococcus macedonicus* ACA-DC 198 points towards the dairy origin of the species. *PLoS ONE* 10:e0116337. doi: 10.1371/journal.pone.0116337
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066
- Sun, Z., Liu, W., Zhang, J., Yu, J., Zhang, W., Cai, C., et al. (2010). Identification and characterization of the dominant lactobacilli isolated from koumiss in China. *J. Gen. Appl. Microbiol.* 56, 257–265. doi: 10.2323/jgam.56.257
- Thompson, J. (1879). The value of koumiss in wasting diseases. *Br. Med. J.* 1, 270. doi: 10.1136/bmj.1.1466.270-c
- Vincent, S. J., Faber, E. J., Neeser, J. R., Stingele, F., and Kamerling, J. P. (2001). Structure and properties of the exopolysaccharide produced by *Streptococcus macedonicus* Sc136. *Glycobiology* 11, 131–139. doi: 10.1093/glycob/11.2.131
- Wang, L., Christoffersen, C. T., Sorich, M. J., Gerber, J. P., Angley, M. T., and Conlon, M. A. (2013). Increased abundance of *Sutterella* spp. and *Ruminococcus torques* in feces of children with autism spectrum disorder. *Mol. Autism* 4, 42. doi: 10.1186/2040-2392-4-42
- Ward, T. L., Hosid, S., Ioshikhes, I., and Altosaar, I. (2013). Human milk metagenome: a functional capacity analysis. *BMC Microbiol.* 13:116. doi: 10.1186/1471-2180-13-116
- Watanabe, K., Fujimoto, J., Tomii, Y., Sasamoto, M., Makino, H., Kudo, Y., et al. (2009). *Lactobacillus kisonensis* sp. nov., *Lactobacillus otakiensis* sp. nov., *Lactobacillus rapi* sp. nov. and *Lactobacillus sunkii* sp. nov., heterofermentative species isolated from sunki, a traditional Japanese pickle. *Int. J. Syst. Evol. Microbiol.* 59(Pt 4), 754–760. doi: 10.1099/ijss.0.004689-0
- Wu, R., Wang, L., Wang, J., Li, H., Menghe, B., Wu, J., et al. (2009). Isolation and preliminary probiotic selection of lactobacilli from koumiss in Inner Mongolia. *J. Basic Microbiol.* 49, 318–326. doi: 10.1002/jobm.200800047
- Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107
- Zhang, W., Sun, Z., Sun, T., and Zhang, H. (2010a). PCR screening and sequence analysis of iol clusters in *Lactobacillus casei* strains isolated from koumiss. *Folia Microbiol. (Praha)* 55, 603–606. doi: 10.1007/s12223-010-0097-3
- Zhang, W., Yu, D., Sun, Z., Wu, R., Chen, X., Chen, W., et al. (2010b). Complete genome sequence of *Lactobacillus casei* Zhang, a new probiotic strain isolated

- from traditional homemade koumiss in Inner Mongolia, China. *J. Bacteriol.* 192, 5268–5269. doi: 10.1128/JB.00802-10
- Zhang, W., and Zhang, H. (2011). “Fermentation and koumiss,” in *Handbook of Food and Beverage Fermentation Technology*, 2nd Edn, ed. Y. H. Hui (Boca Raton, FL: CRC Press), 165–172.
- Zhang, W. Y., Chen, Y. F., Zhao, W. J., Kwok, L. Y., and Zhang, H. P. (2015). Gene expression of proteolytic system of *Lactobacillus helveticus* H9 during milk fermentation. *Ann. Microbiol.* 65, 1171–1175. doi: 10.3168/jds.2014-8520
- Zhao, W., Chen, Y., Sun, Z., Wang, J., Zhou, Z., Sun, T., et al. (2011). Complete genome sequence of *Lactobacillus helveticus* H10. *J. Bacteriol.* 193, 2666–2667. doi: 10.1128/JB.00166-11

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Yao, Yu, Hou, Hui, Liu, Kwok, Menghe, Sun, Zhang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genotyping by PCR and High-Throughput Sequencing of Commercial Probiotic Products Reveals Composition Biases

Wesley Morovic¹, Ashley A. Hibberd¹, Bryan Zabel¹, Rodolphe Barrangou² and Buffy Stahl^{1*}

¹ Genomics and Microbiome Science, DuPont Nutrition & Health, Madison, WI, USA, ² Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, NC, USA

OPEN ACCESS

Edited by:

Jennifer Ronholm,
Health Canada, Canada

Reviewed by:

Giorgio Giraffa,
Centro di Ricerca per le Produzioni
Foraggere e Lattiero-Casearie
(CREA-FLC), Italy
Young Min Kwon,
University of Arkansas, USA

*Correspondence:

Buffy Stahl
buffy.stahl@dupont.com

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 26 August 2016

Accepted: 19 October 2016

Published: 03 November 2016

Citation:

Morovic W, Hibberd AA, Zabel B, Barrangou R and Stahl B (2016)
Genotyping by PCR and High-Throughput Sequencing of Commercial Probiotic Products Reveals Composition Biases.
Front. Microbiol. 7:1747.
doi: 10.3389/fmicb.2016.01747

Recent advances in microbiome research have brought renewed focus on beneficial bacteria, many of which are available in food and dietary supplements. Although probiotics have historically been defined as microorganisms that convey health benefits when ingested in sufficient viable amounts, this description now includes the stipulation “well defined strains,” encompassing definitive taxonomy for consumer consideration and regulatory oversight. Here, we evaluated 52 commercial dietary supplements covering a range of labeled species using plate counting and targeted genotyping. Strain identities were assessed using methods recently published by the United States Pharmacopeial Convention. We also determined the relative abundance of individual bacteria by high-throughput sequencing (HTS) of the 16S rRNA sequence using paired-end 2 × 250 bp Illumina MiSeq technology. Using these methods, we tested the hypothesis that products do contain the quantitative and qualitative list of labeled microbial species. We found that 17 samples (33%) were below label claim for CFU prior to their expiration dates. A multiplexed-PCR scheme showed that only 30/52 (58%) of the products contained a correctly labeled classification, with issues encompassing incorrect taxonomy, missing species, and un-labeled species. The HTS revealed that many blended products consisted predominantly of *Lactobacillus acidophilus* and *Bifidobacterium animalis* subsp. *lactis*. These results highlight the need for reliable methods to determine the correct taxonomy and quantify the relative amounts of mixed microbial populations in commercial probiotic products.

Keywords: probiotics, labeling, testing and assessment, *Lactobacillus*, *Bifidobacterium*, multiplex PCR, taxonomy, high-throughput nucleotide sequencing

INTRODUCTION

Whereas microbiology has historically focused on pathogens and infectious agents, recent efforts have established the importance that microbiomes in general and beneficial microbes in particular play in promoting and maintaining human health (Turnbaugh et al., 2007; Human Microbiome Consortium, 2012). The benefits of health-promoting bacteria have fueled several investigations

establishing the genetic and phenotypic basis for probiotic functionalities (Papadimitriou et al., 2015). The International Scientific Association of Probiotics and Prebiotics defines products containing probiotics as those that “deliver live microorganisms with a suitable viable count of well-defined strains with a reasonable expectation of delivering benefits for the well-being of the host” (Hill et al., 2014), expanding the FAO/WHO definition to include strain-level taxonomy.

The global probiotic industry continues to grow and product introductions rely on the conformation to guidance by local regulatory agencies. Specifically required from several different regulatory bodies (Hammett, 2008; Health Canada, 2015) is accurate labeling of consumer products containing live microbials with species-level identity and viability. Probiotic benefits are typically attributed to specific strains, for which safety and efficacy must be established (Branton et al., 2011; Pariza et al., 2015). In some cases, it is necessary to determine and compare the complete genomes of isolates to accurately identify and distinguish particular genotypes using high-resolution nucleic acid analyses, as beneficial metabolic effects are attributed to these key differences in strains (Briczinski et al., 2009; Broadbent et al., 2012; Ruiz-Moyano et al., 2013). Additionally, strains must be present in sufficient viable quantities to confer a probiotic effect, which varies based on consumer and desired effect (Reid et al., 2001; Leyer et al., 2009). The traditional ISO-approved method of determining viable cell count is by serial dilution and selectively culturing cells to result in colony forming units (CFU) per gram or milliliter (International Organization for Standardization, 2003). While product packaging provides the CFU level content, these are most often reported as the total CFU of a dose and not of individual species or strains. Furthermore, CFU are sometimes measured at time of manufacture although they are known to decrease over time depending on environmental stressors and strain characteristics (Sanders et al., 2014). Indeed, maintaining viability over the course of storage is a major challenge and focus for the probiotics industry.

Advances in sequencing technologies, assembly, and annotation have enabled the scientific community to determine the complete genomes of probiotic strains (Altermann et al., 2005; Stahl and Barrangou, 2013), allowing the development of genotyping methods (Barrangou et al., 2009; Barrangou and Horvath, 2012), and providing unequivocal insights into the proper taxonomy of broadly used commercial strains (Makarova et al., 2006; Briczinski et al., 2009; Loquasto et al., 2013; Milani et al., 2013; Holzapfel and Wood, 2014; Lugli et al., 2014). Trends regarding the formulation of increasingly efficacious and complex blends of multiple probiotics in food and dietary supplements demand the development of high-resolution, yet affordable methods that enable the determination of bacterial counts, and their classification for proper labeling. Some surveys of commercial probiotics have been reported previously, in which authors tested congruence with label claim for phylogenetic identity and CFU counts (Lewis et al., 2015; Patro et al., 2016). Several reports analyzing probiotic claims at the species level focus on ribosomal-based methods. Arguably

the gold-standard in prokaryotic taxonomic identification, the 16S rRNA gene contains homologous and polymorphic sequence regions that can be leveraged in techniques including PCR (Angelakis et al., 2011) and subsequent restriction digest banding (Moreira et al., 2005) to affirm taxonomic classification. By combining 16S rRNA gene PCR with High-Throughput Sequencing (HTS) techniques, the relative abundance of bacteria in a sample can be examined in the form of sequence reads (Caporaso et al., 2011). Indeed, this strategy has been revolutionary in assessing microbiomes in many sample types (Cho and Blaser, 2012; De Leoz et al., 2015; Forssten et al., 2015; Butteiger et al., 2016). There are, however, known limitations to using 16S rRNA sequences, including the presence of multiple heterogeneous copies within a single genome and high sequence similarity between species and sub-species (Dahllöf et al., 2000; Mohkam et al., 2016). This is a known challenge for probiotic genera, notably *Bifidobacterium* and *Lactobacillus* (Milani et al., 2014; Sun et al., 2015). In addition to ribosomal genes, whole genome sequences reveal many other conserved genes that offer higher resolution genotyping opportunities (Figure 1). One such gene is *glucose-6-phosphate isomerase* (*pgi*; EC: 5.3.1.9), a single copy gene whose enzyme catalyzes the important reversible reaction of D-glucose-6-phosphate to D-fructose-6-phosphate in the pentose phosphate pathway and glycolysis (Kanehisa et al., 2016). The aforementioned pathways are conserved biochemical cornerstones of most bacteria, and can actually serve as phylogenetic biomarkers (Brandt and Barrangou, 2016). Furthermore, these genes are widespread, well-annotated, and can be leveraged for taxonomic applications.

Typically, the genus and species binomial nomenclature, together with a total or species-attributed CFU count, are reported on the label of probiotic products. Most probiotic dietary supplement products contain a blend of strains representing various combinations of bacterial genera and species, occasionally including a particularly well-documented strain, formulated with various additional ingredients depending on the delivery format. We surveyed a large set of commercial probiotic samples ($n = 52$) to test whether products meet or exceed the labeled amount, quantitatively, and determine if they are properly labeled, qualitatively. We hypothesized that probiotic blends are formulated to have several key strains that over-represent the total CFU, while other strains are present at lower quantities. Testing the overall viable count claim on labels was completed using traditional plating and species/sub-species identity was surveyed using a novel multiplex PCR (mPCR) targeting polymorphism within the *pgi* gene. Some products also listed strain designations, which we assessed using strain-specific methods (United States Pharmacopeial Convention, 2015) for seven commonly used probiotics. We then used 16S rRNA PCR and HTS to evaluate the relative abundance of species within product formulations. Our results show that there are a select few key probiotics that account for the majority of probiotic blends and high-resolution molecular testing must replace general bacterial surveys to determine qualitative and quantitative contents of probiotic products.

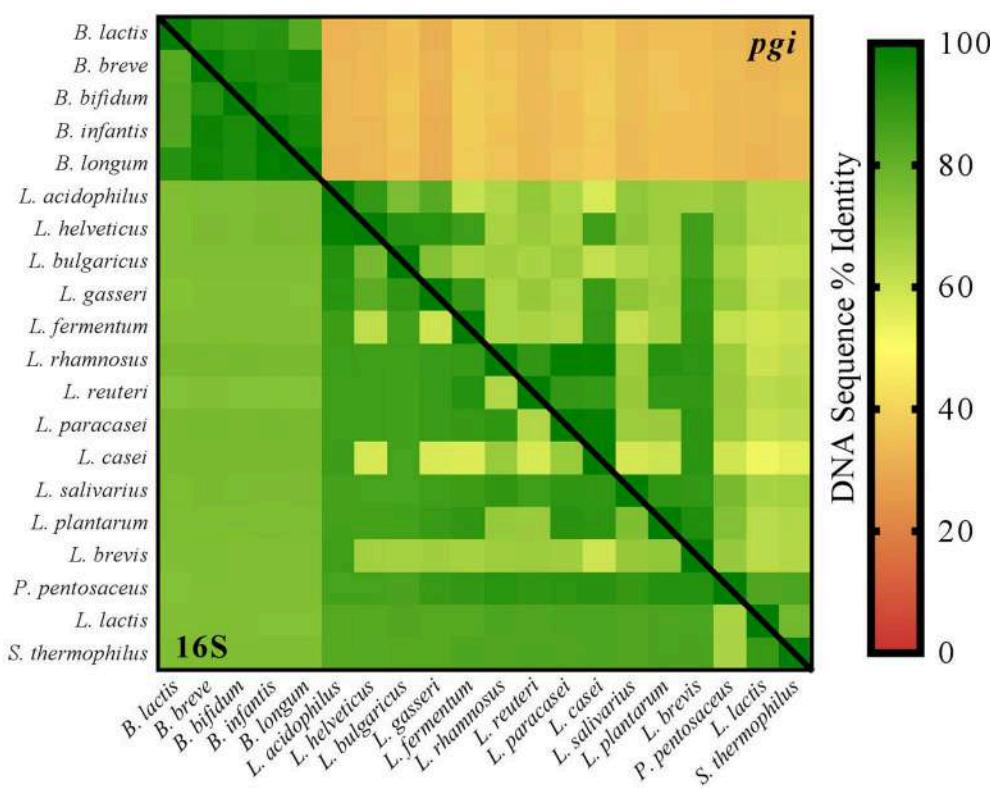


FIGURE 1 | Sequence homology of the 16S rRNA and the glucose-6-phosphate isomerase sequences in commercial probiotics. Gene sequences for the 16S rRNA and *pgf* genes were aligned separately for all 20 organisms listed using the Geneious alignment algorithm. The resulting percent identity matrices were combined into one table and visualized by heat map using Prism 7.01. The overall sequence similarity is higher in the 16S rRNA genes than the *pgf* genes, which presents opportunities for higher resolution assays.

MATERIALS AND METHODS

Product and Standard Preparation

Commercial probiotics were purchased from several retailers in Madison, WI and stored at 4°C within the end shelf-life to decrease cellular mortality. Product contents, including bacteria species, potency, capsule materials, enzymes, flavorings, and other ingredients were noted (Table S1). Samples were resuspended 1:10 (*w/v*) aseptically, remaining encapsulated when possible, and added to 1X Tris-EDTA, pH 8.0 (1X TE) (ThermoFisher p/n BP2473-1). One sample contained chocolate and was resuspended 1:100 (*w/v*) in 1X TE buffer and then incubated in a water bath at 37°C for 30 min to melt. All samples were vortexed until homogenized before further manipulation.

Samples of single-strain freeze-dried concentrates representing all 20 of the species and sub-species in the mPCR were obtained from DuPont and similarly weighed and diluted 1:10 (*w/v*) in 1X TE buffer to serve as standards for validation. The standards were previously tested with the ISO method for CFU and partial 16S rRNA sequence for species identity. Standard samples were created by combining concentrates prior to genomic DNA (gDNA) extraction (labeled with _CFU) as well as extracting each concentrate separately and then combining (labeled with _DNA), using 1X TE buffer for all dilutions. Three

subsets of standards were created for each set: one with template from all samples at equal CFU (all), four standards with only the organisms in each reaction in equal amounts (rxnA-D, see Section mPCR Primer Design), and three mock communities with over-represented *Lactobacillus rhamnosus*, *Lactobacillus acidophilus*, and *Bifidobacterium animalis* subsp. *lactis* (Lrha, Laci, Blac). Final sample concentrations of _CFU standards before gDNA extraction were: all_CFU: 1×10^8 CFU/mL of the 20 targets; rxnA-D_CFU: 1×10^8 CFU/mL of the targets in each reaction; and Lrha_CFU, Laci_CFU, Blac_CFU: 1×10^8 CFU/mL of key targets, 1×10^5 CFU/mL of the other 19 organisms. The final concentrations of the DNA standards were: all_DNA: 100 pg/μL of the 20 targets; rxnA-D_DNA: 1 ng/μL of the five targets in each reaction; and Lrha_DNA, Laci_DNA, Blac_DNA: 1 ng/μL of key target, 1 pg/μL of the other 19 organisms. Standard concentrations are defined further in Table S6.

Genomic DNA Preparation

gDNA was extracted from 250 μL of the 1:10 dilutions of all samples and standards as described above using the MoBio Powersoil gDNA Extraction Kit (MoBio Laboratories, Carlsbad, CA) according to the manufacturer's protocol. Negative controls were included to prevent contamination

of upstream testing. Individual standard gDNA was analyzed by Nanodrop spectrophotometry (ND-1000, Nanodrop, Wilmington, DE) for purity and Qubit (Qubit 2.0, Life Technologies, Carlsbad, CA) for concentration. Standards were further analyzed by electrophoresis of 2% (*w/v*) agarose gel (p/n 17852, ThermoFisher) in 1X Tris-acetate-EDTA (p/n B49, ThermoFisher) stained for 15 min with 1% (*w/v*) Ethidium bromide (p/n E-8751, Sigma) in DI water and de-stained in DI water for 15 min before visualization with UV light (Gel Logic 1500, Kodak, Rochester, NY). All gDNA was stored in -20°C until use. Statistical tests were performed using Minitab 17 (Minitab, State College, PA) and Prism 7.01 (GraphPad, La Jolla, CA). Figures were made using Prism and Geneious v. 6.1.8 (Biomatters Ltd., Auckland, New Zealand).

Assessment of Total Colony Forming Units

Samples were resuspended 10% (*w/v*) in buffered peptone water (p/n FTPW9966, 3M) and serially diluted sufficiently to test the label claim of CFU. A pour plate technique was used, where 1 mL of the final dilution and 15 mL of deMan, Rogosa, and Sharpe (MRS) agar (p/n 288210, BD Difco, Franklin Lakes, New Jersey) supplemented with 0.05% cysteine-HCl (p/n C7880, Sigma, St. Louis, MO) were added to three replicate petri dishes. The plates were swirled gently to homogenize and cooled at room temperature until the agar solidified. Plates were incubated anaerobically at 37°C for 48 h. Resulting colonies were multiplied by the dilution factor and averaged between the replicates to give final CFU/g. This method provides enrichment for all 20 of the organisms in the following assays.

mPCR Primer Design

Complete *pgi* sequences (Table S2) were extracted from both non-public DuPont culture collection genomes and those in the National Centre for Biotechnology (NCBI, Bethesda, MD) and the Genomes Online Database (JGI, Walnut Creek, CA). Sequences were categorized by species or sub-species based upon whole genome alignments and full-length 16S identity. Pairwise alignment was then performed using Geneious alignment algorithm and resulted in a consensus sequence with degenerate nucleotides representing 100% sequence identity (Figures S1A,B). All alignments were made using the default input values. The consensus sequences were then compared to the top 100 matches using the Basic Local Alignment Search Tool (*blastn*; NCBI) to locate suitable priming targets compared to the closest related sequences. Primers were designed for each species or sub-species (Table S3) and then tested for hairpins and dimers using OligoAnalyzer (Integrated DNA Technologies, Coralville, IA). Further *in silico* analysis was performed using *blastn* to prevent possible amplification of undesired targets. The assays were grouped into four pentaplex reactions (rxnA-D) based on amplicon length (Figure S2). Oligos were obtained from IDT and rehydrated with 1X TE buffer to a stock concentration of 100 μM and stored at -20°C .

mPCR Validation

Reactions were optimized for primer concentration, annealing temperature, MgCl₂ concentration, dNTPs, and GC enhancer by gradients (data not shown). All primer combinations were

tested individually against all other 19 species and sub-species to assess non-specific primer binding. Primer target specificity was further validated by testing each primer set against up to 10 different strains of each species and sub-species. Limit-of-detection (LOD) and preferential amplification experiments were tested using the standards as listed above. The mPCR reaction formula and thermocycler settings are listed on Table S4.

Probiotic Sample Testing by mPCR and Strain-Specific PCR

Samples and standards were tested according to the PCR procedures listed in Table S4. Amplicons were visualized using 2% agarose gel electrophoresis with ethidium bromide staining as described above or by 2% E-Gel with ethidium bromide (p/n G600002, Invitrogen). Samples requiring sequence confirmation were cleaned with PCR Clean-Up and Gel Extraction Kit (Clontech, Mountain View, CA) and sent to Eurofins Genomics (Eurofins MWG Operon LLC, Louisville, KY) for Sanger sequencing. Sequences were analyzed using Geneious.

High-Throughput 16S rRNA Gene Sequencing

Samples and CFU standards were processed using a custom barcoding scheme as previously described (Caporaso et al., 2011). Briefly, triplicate PCR was performed with the 16S rRNA V4 primers in Table S3 and associated Golay barcodes according to the PCR procedures listed in Table S4. Amplicons were visualized using 2% agarose E-Gels with ethidium bromide, normalized with SequalPrep Normalization Kits (p/n A1051001, Applied Biosystems), pooled and concentrated with Microcon 30K Centrifugal Columns (p/n UFC503024 EMD Millipore, Merck KGaA, Darmstadt, Germany). The amplicon pool was sequenced using 2 \times 250 Paired-End Illumina MiSeq technologies (Pioneer, Johnston, IA) with the addition of 25% PhiX to increase library diversity. Sequencing data was processed using the Quantitative Insights into Microbial Ecology (QIIME v1.9.1) pipeline (Caporaso et al., 2010). Reads were paired using fastq-join (Aronesty, 2011) and filtered to remove reads that contained ambiguous bases or a Phred quality score <30 . The remaining sequences were clustered *de novo* at 100% identity with uclust (Edgar, 2010) and assigned a taxonomic identity using the default Greengenes database (DeSantis et al., 2006; v 13_8) in QIIME. Additionally, taxonomy was manually assigned to *de novo* OTUs representing $>0.1\%$ of the total reads by pairwise alignment to the closest type strain in the EZ-Taxon database (Kim et al., 2012) to achieve greater accuracy and resolution (Figure S4). Phylogenetic trees were generated using Geneious.

RESULTS

Total Viable Count in Samples Compared to Expiration Date

The products listed an array of ingredients including capsule type, excipients, and specialty ingredients such as flavoring, vitamins and minerals, and enzymes such as lactase, lysozyme, and protease (Table S1). The average labeled count was 2.3×10^{10} CFU/g, with minimum and maximum counts of 1.0×10^8

CFU/g and 9.0×10^{10} CFU/g, respectively. Many of the products ($n = 24$) listed the disclaimer that potency measurements on the label were made at the “time of manufacture,” and are represented in red in **Figure 2**. The plated count had an average of 6.6×10^{10} , with a minimum count of 4.7×10^5 and a maximum count of 4.0×10^{11} . Overall, 35 of the samples (67.3%) had total CFU above the label claim, and four of those samples had an excess of over 1 log (**Figure 2A**). Furthermore, products quantified at time of manufacture had significantly less average CFUs than those labeled “to expiration” ($p = 0.015$, 2-Sample *t*-Test) and were trending toward having more samples below the label claim ($p = 0.080$, 2-Sample *t*-Test). The time to expiration from date of measurement was also noted, although there was no correlation between overall time to expiration and congruence with label claim (**Figure 2B**).

mPCR Validation

Primers were validated against multiple strains of each sub-species noted in the assay. The *B. longum* and *B. infantis* assays were compared to the BLIR test (Lewis et al., 2015) using gDNA standards and both assays successfully differentiated the sub-species (**Figures S1C–E**). Each primer set was tested against all other species and sub-species standards to confirm no cross-amplification. Some of the higher G+C templates in bifidobacteria did show faint non-specific binding, although amplicon sizes were distinguishable. Furthermore, *in silico* analysis showed that the *B. infantis* and *B. breve* assays may amplify other species of bifidobacteria that are not typically sold as probiotics. All results were considered positive only if the gel bands exactly matched *in silico* amplicon length.

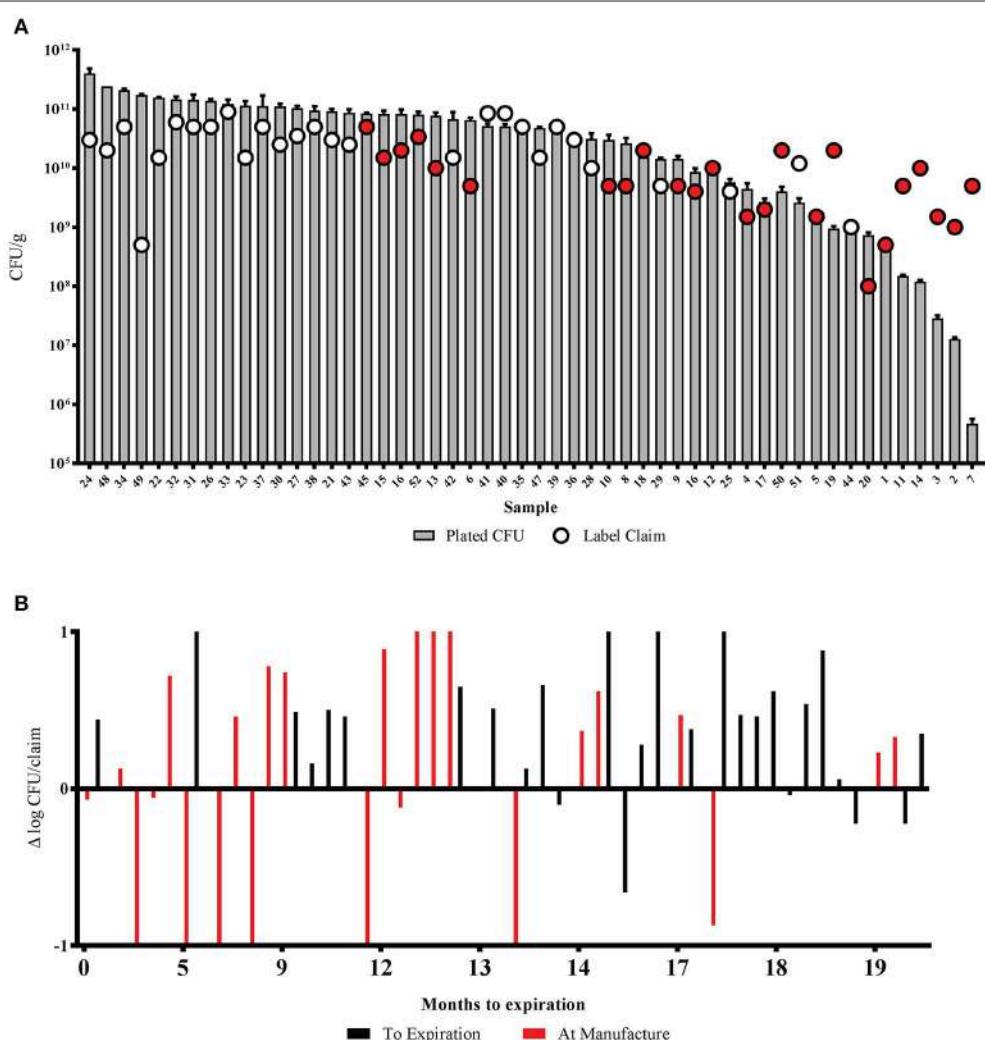


FIGURE 2 | Total colony forming units of probiotic products compared to labeled potency. (A) The CFU of each sample compared to the label claim.

Samples are organized by decreasing total CFU/g. Error bars show the standard deviation of each triplicate plate count. **(B)** The months until expiration is noted on the horizontal axis. Samples above the 0 y-axis gridline are above label claim, and those below are below label claim. All samples in red claimed potency at the time of manufacture.

Measurement of the individual standard gDNA extractions showed that on average, 1.2×10^{10} cells produced a yield of 51.8 ng/ μ L of high molecular weight DNA with an average A280/A260 nm of 1.93. One sample had 2.1×10^8 CFU and had a gDNA yield of 11.6 ng/ μ L. No significant sampling yield bias was seen between the lactobacilli and bifidobacteria standards (One-way ANOVA, $p = 0.261$). The all_CFU controls amplified in all reactions, showing that there is no inhibition of individual reactions when all are blended at similar concentrations. The all_DNA control was serially diluted to 1 pg/ μ L and all but one reaction amplified. This adheres to the definition of LOD as the lowest concentration at which 95% of positive samples are detected (Bustin et al., 2009). The Lrha_CFU, Laci_CFU, Blac_CFU controls all amplified the labeled species, however only 78.9% of the lower dilutions amplified for the _CFU samples and none of the _DNA lower dilution targets amplified. This shows that high concentrations of single organisms can have an inhibitory effect on targets with lower concentrations. Considering the above controls, the mPCR assay is effective for blends of the listed target bacteria that are at 1 pg/ μ L, or 2.3×10^5 cells of starting material, which is below most recommended effective probiotic doses.

Accuracy of Bacterial Species Labeling in Probiotic Products

Probiotic samples were tested and assessed by comparing resulting amplicons to positive control ladders (Figure S2). Discrepancies from the label claim were retested to rule out PCR error. All amplicons were compared further against the 16S profiling and all results were mapped based on the label claims (Figure 3A). Overall, 11 (21.1%) of the blended products with at least two organisms had one or more claimed probiotic organisms missing or too diluted to detect. Conversely, 18 (34.6%) blended products had an additional organism not listed on the label claim. Considering some samples had both unlabeled positives and labeled negatives, 22 (42.3%) of the samples tested showed evidence of having incorrectly listed the target species and sub-species, only two of which did not originally claim one of the target genus *Bifidobacterium*. Some samples appeared to have switched some species, such as *B. infantis* for *B. longum* (29 and 30), *L. paracasei* for *L. casei* (29, 30, and 50), and *L. helveticus* for *L. acidophilus* (32, 51 and 52). Noteworthy, some of these closely related species have been historically difficult to distinguish until recent advances in molecular biology.

Detecting Strains in Blended Products

In general, 18 (34.6%) of the products listed specific strain designations. Samples that matched the species of the target strains were tested with the strain-specific primers listed in Table S3 and are visualized in Figure 3B. Two of the samples (33 and 44) were confirmed to have incorrect *L. acidophilus* strains labeled. Four samples (33, 42–44) incorrectly labeled the presence of *L. acidophilus* strain NCFM, where it was not found to be present. The majority of products (36/41) that contain an *L. acidophilus* were confirmed to have strain La-14. Two samples (22 and 43) had multiple *B. lactis* strains indistinguishable by SNP typing. Some products listed strains that do not have USP

reference methods and therefore their strain designation could not be confirmed.

Relative Abundance of Organisms in the Product Microbial Blends

We focused analysis on OTUs comprising more than 0.1% of total sequencing reads which resulted in 42 OTUs that were grouped to type strains using EzTaxon (Table S5, Kim et al., 2012). Most species were individually distinguishable except for several closely related species: the *B. breve* group that included *B. longum* and *B. infantis*; the *L. casei* group that included *L. paracasei*; and the *L. acidophilus* group that included *L. helveticus*. The average reads per sample after quality filtering was 58,144 reads and only one sample (14) had <10,000 reads and was removed from analysis. Controls were also sequenced to assess the accuracy of the abundance calculations (Figure 4A). Comparisons of standard CFU dilutions to percent reads resulted in a linear regression line with an R-squared value of 89.6% (simple regression). HTS results were ordered based on the two most abundant species overall in all tested products, namely *L. acidophilus* and *B. lactis*, which represented 35.6 and 15.7% of all sample reads, respectfully (Figure 4B, Figure S3). *L. acidophilus* was significantly more abundant than all other species, even after removing samples positive for *L. helveticus* using the mPCR, while *B. lactis* was significantly more abundant than all species but *B. breve*, *L. plantarum*, *L. rhamnosus*, *L. gasseri*, and *L. reuteri* ($p < 0.05$, Tukey's multiple comparisons test). Furthermore, the two species on average made up 65.8% of the reads in blends with at least two probiotics. The abundances of the top 10 probiotics in each product positive for 10 or more probiotics with the mPCR test were assessed. The average abundance of the top OTU group in each product by input was nearly two orders of magnitude higher than the 10th ranked OTU group (40.7–0.6%, respectfully), and the general decrease in product abundance fits a decreasing logarithmic curve (R -squared 92.6%, simple regression; Figure 4C).

DISCUSSION

The strain-specific health benefits and safety of probiotics are of utmost importance to dietary supplement industry leaders, researchers, regulatory entities, and consumer groups. Current identification methods are not amenable to mixed microbial communities and therefore probiotic bacteria must be correctly identified prior to blending of multiple strains across genera and species. Although efficacy research on probiotics is increasing, without correct identification and labeling strains cannot be associated with specific studies. Additionally, this makes it difficult for consumers to choose products based on health claims associated with specific strains. With increasing focus on the small genomic variations that differentiate strains, and the potential metabolic repercussions therein, there is a great need to use high-resolution genotyping methods to assure identity as well as quantity. Recently, the International Probiotic Association approved Flow Cytometry as a method

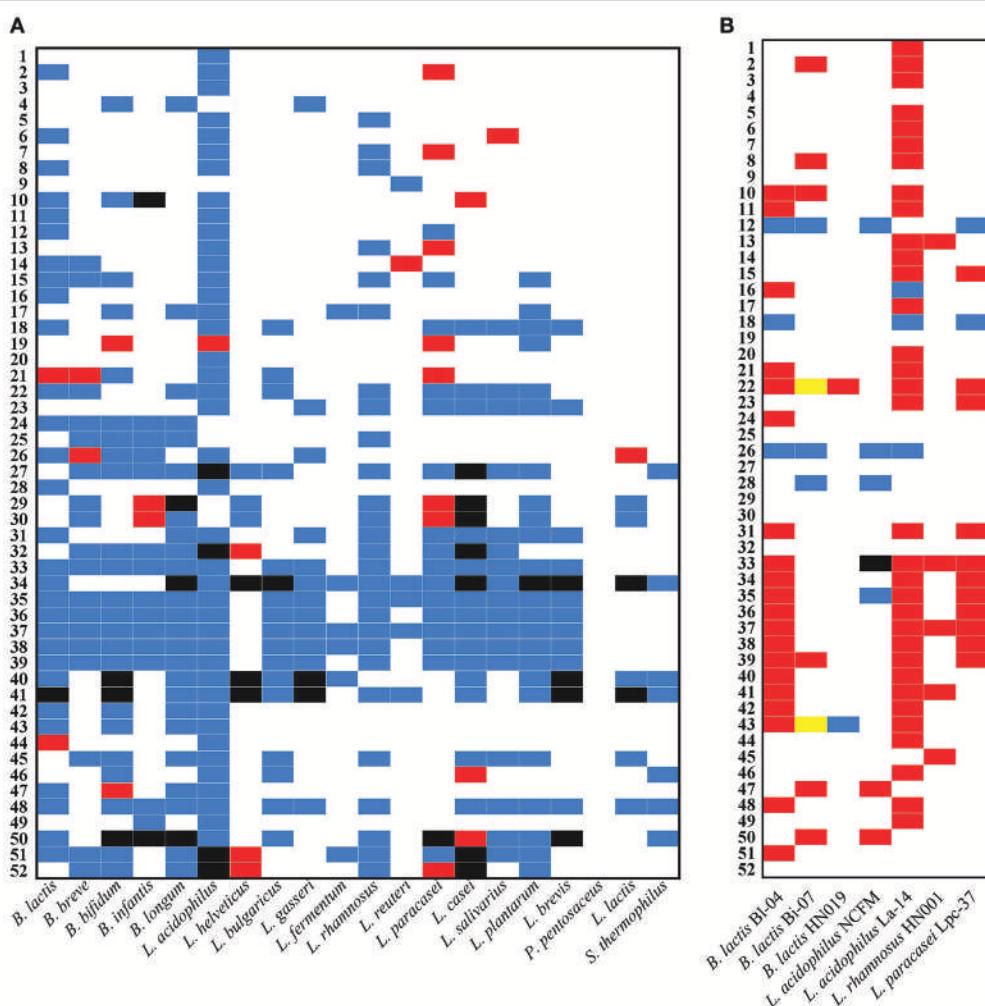


FIGURE 3 | PCR assay of species, sub-species, and strain identity compared to label claim. The presence of organisms is visualized for (A) the mPCR and (B) the strain-specific PCR assays as different colors: blue denotes claimed and present; white denotes not claimed and absent; red denotes not claimed and present; black denotes claimed and not present; yellow denotes strain-specific assays unable to be fully characterized using the present assays. Samples are ordered based on species taxonomy by 16S rRNA sequence.

for enumerating probiotic bacteria (International Organization for Standardization, 2015). This method quickly detects viable cells to satisfy the traditional definition of probiotics as “live microorganisms which, when administered in adequate amounts, confer a health benefit on the host” (FAO-WHO, 2001), but does not differentiate cells based on genotype (Davis, 2014). Colony morphologies of closely related species and sub-species will not suffice for distinguishing the very closely related probiotic bacteria, making DNA-based methods the best option for classification.

Several other molecular methods have been evaluated for probiotic testing, including qPCR (Postollec et al., 2011) and microarrays (Patro et al., 2015). While many of these reports have identified incorrect labeling of probiotic organisms, no single test has been proposed to survey mixed microbial finished goods using a single gene target. One report from the FDA recognized

this technology gap and introduced a test based on shotgun sequencing, and utilizing a custom in-house bioinformatics pipeline (Patro et al., 2016). While metagenomic sequencing can offer the resolution needed to identify all strains within a sample, simpler methods like PCR provide more resolution than current standards, with less expense and time to release. An industry accepted intermediate method must affordably, rapidly, and accurately provide species-level resolution regardless of product formulation.

In this study, we sought to understand the baseline of regulatory compliance by investigating label claims of 52 commercial probiotic products for quantity and genetic identity. Culture-based plating methods showed that a majority of products ($n = 35$) contained total amounts of viable probiotics above the label claim, which is higher than a previous report (Weese and Martin, 2011). Perhaps not surprisingly, products

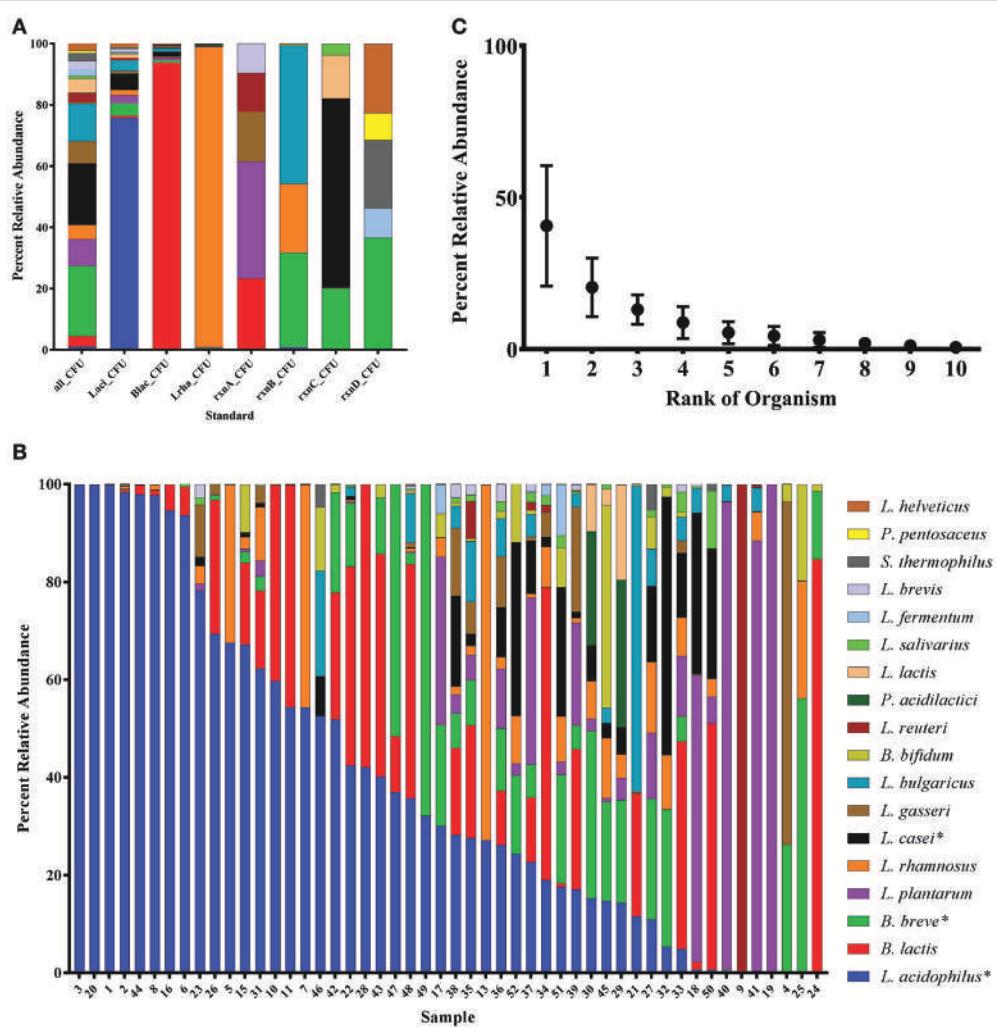


FIGURE 4 | High-throughput sequencing of the 16S rRNA gene in probiotic products. Bar graphs show the percent abundance of **(A)** controls and **(B)** commercial samples. Filtered reads are not shown. OTUs that represent more than one organism have asterisks by the species name. Samples are ordered by decreasing key species: first *L. acidophilus*, then *B. lactis*, and *B. breve* abundances. **(C)** Thirteen of the samples had 10 or more probiotics detected using the mPCR assay. Each OTU in the samples was ranked from high to low abundance regardless of identity, and the averages are noted as dot plots with error bars representing the standard deviation.

quantified at time of manufacture showed less overall CFU, and many were below label claim well before the listed expiration date. Our novel mPCR assay enabled relatively rapid detection of 20 distinct probiotic species and sub-species. The authors acknowledge that as new species of probiotics are introduced, they may or may not contain *pgi* genes and that these additions will need to be designed and validated to flexibly fit with the method described herein. This new method revealed identification discrepancies for 22 of the 52 products, several of which were likely due to misidentification of sub-species. While labels may be technically correct in identifying a species, it is important to denote the correct sub-species. For example, *B. infantis* is often used as a dietary supplement to establish infant microbiota in the presence of human milk oligosaccharides, a function that has not been

demonstrated by *B. longum* (LoCascio et al., 2010). Some labels acknowledged incorrect classification, such as sample 49 that read “*B. infantis* (*B. lactis*)” which is scientifically incorrect and likely confusing to consumers. While few products listed the strain content on the label, the strain-specific USP testing demonstrated that it is possible to identify highly clonal strains using traditional PCR methods, and is verifiable when indicated. Finally, the HTS based on 16S rRNA gene sequence showed an uneven abundance of probiotics in blended products, where the most dominant strains, particularly *L. acidophilus* and *B. lactis*, represented over half of the reads in all of the samples.

These results clearly show probiotic strains in these dietary supplements were characteristically not of equal distribution, similar to results demonstrated previously by HTS (Patro

et al., 2015) and microarray analyses (Angelakis et al., 2011). Although the CFU testing only determined viable cells, PCR can successfully amplify intact extracellular DNA or DNA from dead cells (Kramer et al., 2009), so the abundance of viable cells for the different probiotics could be different. As mentioned, many different methods are available to determine viable amounts of specific species (Davis, 2014), although it still remains difficult to quantify mixed microbial constituents in routine industrial and commercial product quality assessment. While this study does not provide an ideal solution to measure viability of different strains, it does highlight a gap in methods to determine the shelf-life stability of individual strains after they have been combined into a commercial mixture.

Although not demonstrated here, we hypothesize that manufacturers are perhaps formulating products based on the stability or cost of particular strains, or even on consumer awareness of select species. While formulations of input strains seemed skewed for high abundance of a few species, sequencing also demonstrated that the ingredient strains seem to be free of any other microbial contamination, including any pathogenic species, filtered at 0.1%. While not comprehensive, 46 of the commercial probiotic products we surveyed included more than one organism, further highlighting the need for a technique to determine each of the major input organisms at the species level.

Having established that a significant proportion of commercial probiotic products do not meet basic requirements of the correct taxonomic group (mostly at the species level) listed on the ingredients list, we developed methods that enable the industry to identify and release probiotic products. These methods will also help formulate, blend, and label probiotic products to meet the necessary standards for the regulatory agencies and consumer groups alike. As the health-promoting roles of bacteria become more substantiated, and the biochemical functions attributed to microbiomes advance toward therapeutic applications, it will be paramount to use sound, state-of-the-art, and affordable methods to formulate commercial products and document their composition.

DATA DEPOSITION

The HTS data has been deposited in the Sequence Read Archive in NCBI under accession number SRP090599 and in Qiita as ID 10681. Proprietary *pgi* gene sequences have been uploaded to GenBank in NCBI and can be accessed using the accession numbers listed in **Table S2**.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: WM, RB, and BS. Performed the experiments: WM, AH, and BZ. Contributed

analysis: WM, AH, RB, and BS. Wrote the paper: WM, RB, AH, and BS.

FUNDING

RB and the Klaenhammer Laboratory at North Carolina State University are supported by a research contract from DuPont Nutrition & Health.

ACKNOWLEDGMENTS

The authors thank Martha DeMeules for performing the CFU measurements and Paige Roos at the DuPont Pioneer Genomics Sequencing Facility for library construction and HTS. We also thank Sarah Hansen and Steve Prescott for thoughtful discussion.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2016.01747/full#supplementary-material>

Figure S1 | Multiplex assay design. Alignments of the *pgi* genes are shown for **(A)** lactobacilli and **(B)** bifidobacteria. Primers are shown as arrows above the target sequence. Black lines in the sequence blocks denote polymorphism. Specific examples of primers are shown with **(C)** *B. longum* reverse, **(D)** *B. infantis* forward, and **(E)** *B. infantis* reverse. Strains with (T) represent the type strain of each sub-species.

Figure S2 | Visualization of the positive bands from the mPCR assay. A virtual gel created in Geneious is next to an E-Gel® with the results of testing gDNA standards. The band identities and amplicon sizes are listed to the right.

Figure S3 | Heat map of relative abundance of probiotics in samples based on high-throughput sequencing. The relative abundance of probiotics in each sample is visualized with shaded cells, with darker shading representing higher abundance. Blue cells represent the presence of organisms that are claimed by the products, while red cells are organisms not claimed by products. OTUs that represent more than one organism have asterisks by the species name.

Figure S4 | Mapping of the OTUs to type strains in EzTaxon. The alignment tree that maps each of the 42 OTUs to 16S rRNA gene sequences from type strains of each of the 20 target species. Tree was generated using Geneious Tree Builder with default settings.

Table S1 | Sample contents. Gray boxes denote a probiotic claimed in the product.

Table S2 | Strains used to develop the mPCR *in silico*. Accession numbers for whole genome sequences are noted if available. GenBank accession numbers are noted for strains without associated genome sequences.

Table S3 | Primers used in the study.

Table S4 | Settings used for all PCR assays.

Table S5 | HTS read analysis.

Table S6 | Standard concentrations.

Angelakis, E., Million, M., Henry, M., and Raoult, D. (2011). Rapid and accurate bacterial identification in probiotics and yogurts by MALDI-TOF mass spectrometry. *J. Food Sci.* 76, M568–M572. doi: 10.1111/j.1750-3841.2011.02369.x

Aronesty, E. (2011). *Ea-Utils: Command-Line Tools for Processing Biological Sequencing Data. Expression Analysis*. Durham, NC.

REFERENCES

- Altermann, E., Russell, W. M., Azcarate-Peril, M. A., Barrangou, R., Buck, B. L., McAuliffe, O., et al. (2005). Complete genome sequence of the probiotic lactic acid bacterium *Lactobacillus acidophilus* NCFM. *Proc. Natl. Acad. Sci. U.S.A.* 102, 3906–3912. doi: 10.1073/pnas.0409188102

- Barrangou, R., Brzczinski, E. P., Traeger, L. L., Loquasto, J. R., Richards, M., Horvath, P., et al. (2009). Comparison of the complete genome sequences of *Bifidobacterium animalis* subsp. *lactis* DSM 10140 and Bl-04. *J. Bacteriol.* 191, 4144–4151. doi: 10.1128/JB.00155-09
- Barrangou, R., and Horvath, P. (2012). CRISPR: new horizons in phage resistance and strain identification. *Annu. Rev. Food. Sci. Technol.* 3, 143–162. doi: 10.1146/annurev-food-022811-101134
- Brandt, K., and Barrangou, R. (2016). Phylogenetic analysis of the *Bifidobacterium* genus using glycolysis enzyme sequences. *Front. Microbiol.* 7:657. doi: 10.3389/fmicb.2016.00657
- Branton, W., Jones, M., Tomaro-Duchesneau, C., Martoni, C., and Prakash, S. (2011). *In vitro* characterization and safety of the probiotic strain *Lactobacillus reuteri* cardioviva NCIMB 30242. *Int. J. Probiotics Prebiotics* 6, 1–12.
- Brzczinski, E. P., Loquasto, J. R., Barrangou, R., Dudley, E. G., Roberts, A. M., and Roberts, R. F. (2009). Strain-specific genotyping of *Bifidobacterium animalis* subsp. *lactis* by using single-nucleotide polymorphisms, insertions, and deletions. *Appl. Environ. Microbiol.* 75, 7501–7508. doi: 10.1128/AEM.01430-09
- Broadbent, J. R., Neeno-Eckwall, E. C., Stahl, B., Tandee, K., Cai, H., Morovic, W., et al. (2012). Analysis of the *Lactobacillus casei* supragenome and its influence in species evolution and lifestyle adaptation. *BMC Genomics* 13:533. doi: 10.1186/1471-2164-13-533
- Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., et al. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.* 55, 611–622. doi: 10.1373/clinchem.2008.112797
- Butteiger, D. N., Hibberd, A. A., McGraw, N. J., Napawan, N., Hall-Porter, J. M., and Krul, E. S. (2016). Soy protein compared with milk protein in a western diet increases gut microbial diversity and reduces serum lipids in golden syrian hamsters. *J. Nutr.* 146, 697–705. doi: 10.3945/jn.115.224196
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4516–4522. doi: 10.1073/pnas.1000080108
- Cho, I., and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* 13, 260–270. doi: 10.1038/nrg3182
- Dahllöf, I., Baillie, H., and Kjelleberg, S. (2000). rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. *Appl. Environ. Microbiol.* 66, 3376–3380. doi: 10.1128/AEM.66.8.3376-3380.2000
- Davis, C. (2014). Enumeration of probiotic strains: review of culture-dependent and alternative techniques to quantify viable bacteria. *J. Microbiol. Methods* 103, 9–17. doi: 10.1016/j.mimet.2014.04.012
- De Leo, M. L., Kalanetra, K. M., Bokulich, N. A., Strum, J. S., Underwood, M. A., German, J. B., et al. (2015). Human milk glycomics and gut microbial genomics in infant feces show a correlation between human milk oligosaccharides and gut microbiota: a proof-of-concept study. *J. Proteome Res.* 14, 491–502. doi: 10.1021/pr500759e
- DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461
- FAO-WHO (2001). *Health and Nutritional Properties of Probiotics in Food Including Powder Milk with live Lactic Acid Bacteria*. Geneva: World Health Organization. Available at: http://www.who.int/foodsafety/publications/fs_management/en/probiotics.pdf (Accessed August 26, 2016).
- Forssten, S. D., Röytöö, H., Hibberd, A. A., and Ouwehand, A. C. (2015). The effect of polydextrose and probiotic lactobacilli in a *Clostridium difficile*-infected human colonic model. *Microb. Ecol. Health. Dis.* 26:27988. doi: 10.3402/mehd.v26.27988
- Hammett, R. (2008). *Therapeutic Goods Order No. 78: Standard for Tablets and Capsules*. Australian Government Therapeutic Goods Administration, F2008L04287. Available online at: <https://www.legislation.gov.au/Details/F2008L04287> (Accessed August 26, 2016).
- Health Canada (2015). *Natural Health Product: Probiotics*. Health Canada. Available online at: <http://webprod.hc-sc.gc.ca/nhpid-bdipsn/atReq.do?atid=probio&lang=eng> (Accessed August 26, 2016).
- Hill, C., Guarner, F., Reid, G., Gibson, G. R., Merenstein, D. J., Pot, B., et al. (2014). Expert consensus document, the international scientific association for probiotics and prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nat. Rev. Gastroenterol. Hepatol.* 11, 506–514. doi: 10.1038/nrgastro.2014.66
- Holzapfel, W. H., and Wood, B. J. (eds.). (2014). *Lactic Acid Bacteria: Biodiversity and Taxonomy*. Wiley Blackwell.
- Human Microbiome Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234
- International Organization for Standardization (2003). ISO 7889:2003 (IDF 117:2003). *Yogurt — Enumeration of Characteristic Microorganisms — Colony-Count Technique at 37 Degrees C*. International Organization for Standardization. Available online at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=31880 (Accessed August 26, 2016).
- International Organization for Standardization (2015). ISO 19344:2015 (IDF 232): *Milk and Milk Products – Starter Cultures, Probiotics and Fermented Products – Quantification of Lactic Acid Bacteria by Flow Cytometry*. International Organization for Standardization. Available online at: http://www.iso.org/iso/catalogue_detail?csnumber=64658 (Accessed August 26, 2016).
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi: 10.1093/nar/gkv1070
- Kim, O. S., Cho, Y. J., Lee, K., Yoon, S. H., Kim, M., Na, H., et al. (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int. J. Syst. Evol. Microbiol.* 62, 716–721. doi: 10.1093/ijss.038075-0
- Kramer, M., Obermajer, N., Bogovic Matijasic, B., Rogelj, I., and Kmetec, V. (2009). Quantification of live and dead probiotic bacteria in lyophilised product by real-time PCR and by flow cytometry. *Appl. Microbiol. Biotechnol.* 84, 1137–1147. doi: 10.1007/s00253-009-2068-7
- Lewis, Z. T., Shani, G., Masarweh, C. F., Popovic, M., Frese, S. A., Sela, D. A., et al. (2015). Validating bifidobacterial species and subspecies identity in commercial probiotic products. *Pediatr. Res.* 79, 445–452. doi: 10.1038/pr.2015.244
- Leyer, G. J., Li, S., Mubasher, M. E., Reifer, C., and Ouwehand, A. C. (2009). Probiotic effects on cold and influenza-like symptom incidence and duration in children. *Pediatrics* 124, e172–e179. doi: 10.1542/peds.2008-2666
- LoCascio, R. G., Desai, P., Sela, D. A., Weimer, B., and Mills, D. A. (2010). Broad conservation of milk utilization genes in *Bifidobacterium longum* subsp. *infantis* as revealed by comparative genomic hybridization. *Appl. Environ. Microbiol.* 76, 7373–7381. doi: 10.1128/AEM.00675-10
- Loquasto, J. R., Barrangou, R., Dudley, E. G., Stahl, B., Chen, C., and Roberts, R. F. (2013). *Bifidobacterium animalis* subsp. *lactis* ATCC 27673 is a genetically unique strain within its conserved subspecies. *Appl. Environ. Microbiol.* 79, 6903–6910. doi: 10.1128/AEM.01777-13
- Lugli, G. A., Milani, C., Turroni, F., Duranti, S., Ferrario, C., Viappiani, A., et al. (2014). Comparative genomics to investigate the evolutionary development of the genus *Bifidobacterium*. *Appl. Environ. Microbiol.* 80, 6383–6394. doi: 10.1128/AEM.02004-14
- Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., et al. (2006). Comparative genomics of the lactic acid bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 103, 15611–15616. doi: 10.1073/pnas.0607117103
- Milani, C., Duranti, S., Lugli, G. A., Bottacini, F., Strati, F., Arioli, S., et al. (2013). Comparative genomics of *Bifidobacterium animalis* subsp. *lactis* reveals a strict monophyletic bifidobacterial taxon. *Appl. Environ. Microbiol.* 79, 4304–4315. doi: 10.1128/AEM.00984-13
- Milani, C., Lugli, G. A., Duranti, S., Turroni, F., Bottacini, F., Mangifesta, M., et al. (2014). Genomic encyclopedia of type strains of the genus *Bifidobacterium*. *Appl. Environ. Microbiol.* 80, 6290–6302. doi: 10.1128/AEM.02308-14
- Mohkam, M., Nezafat, N., Berenjian, A., Mobasher, M. A., and Ghasemi, Y. (2016). Identification of *Bacillus* Probiotics Isolated from Soil Rhizosphere Using 16S

- rRNA, recA, rpoB Gene Sequencing and RAPD-PCR. *Probiotics Antimicrob. Proteins* 8, 8–18. doi: 10.1007/s12602-016-9208-z
- Moreira, J. L., Mota, R. M., Horta, M. F., Teixeira, S. M., Neumann, E., Nicoli, J. R., et al. (2005). Identification to the species level of *Lactobacillus* isolated in probiotic prospecting studies of human, animal or food origin by 16S-23S rRNA restriction profiling. *BMC Microbiol.* 5:15. doi: 10.1186/1471-2180-5-15
- Papadimitriou, K., Zoumpopoulou, G., Foligné, B., Alexandraki, V., Kazou, M., Pot, B., et al. (2015). Discovering probiotic microorganisms: *in vitro*, *in vivo*, genetic and omics approaches. *Front. Microbiol.* 6:58. doi: 10.3389/fmicb.2015.00058
- Pariza, M. W., Gillies, K. O., Kraak-Ripple, S. F., Leyer, G., and Smith, A. B. (2015). Determining the safety of microbial cultures for consumption by humans and animals. *Regul. Toxicol. Pharmacol.* 73, 164–171. doi: 10.1016/j.yrtph.2015.07.003
- Patro, J. N., Ramachandran, P., Barnaba, T., Mammel, M. K., Lewis, J. L., and Elkins, C. A. (2016). Culture-independent metagenomic surveillance of commercially available probiotics with high-throughput next-generation sequencing. *mSphere* 1:e00057-16. doi: 10.1128/mSphere.00057-16
- Patro, J. N., Ramachandran, P., Lewis, J. L., Mammel, M. K., Barnaba, T., Pfeiler, E. A., et al. (2015). Development and utility of the FDA ‘GutProbe’ DNA microarray for identification, genotyping and metagenomic analysis of commercially available probiotics. *J. Appl. Microbiol.* 118, 1478–1488. doi: 10.1111/jam.12795
- Postollec, F., Falentin, H., Pavan, S., Combrisson, J., and Sohier, D. (2011). Recent advances in quantitative PCR (qPCR) applications in food microbiology. *Food. Microbiol.* 28, 848–861. doi: 10.1016/j.fm.2011.02.008
- Reid, G., Beuerman, D., Heinemann, C. and Bruce, A. W. (2001). Probiotic *Lactobacillus* dose required to restore and maintain a normal vaginal flora. *FEMS Immunol. Med. Microbiol.* 32, 37–41. doi: 10.1111/j.1574-695X.2001.tb00531.x
- Ruiz-Moyano, S., Totten, S. M., Garrido, D. A., Smilowitz, J. T., German, J. B., Lebrilla, C. B., et al. (2013). Variation in consumption of human milk oligosaccharides by infant gut-associated strains of *Bifidobacterium breve*. *Appl. Environ. Microbiol.* 79, 6040–6049. doi: 10.1128/AEM.01843-13
- Sanders, M. E., Klaenhammer, T. R., Ouwehand, A. C., Pot, B., Johansen, E., Heimbach, J. T., et al. (2014). Effects of genetic, processing, or product formulation changes on efficacy and safety of probiotics. *Ann. N.Y. Acad. Sci.* 1309, 1–18. doi: 10.1111/nyas.12363
- Stahl, B., and Barrangou, R. (2013). Complete genome sequence of probiotic strain *Lactobacillus acidophilus* La-14. *Genome Announc.* 1:e00376-13. doi: 10.1128/genomeA.00376-13
- Sun, Z., Harris, H. M., McCann, A., Guo, C., Argimón, S., Zhang, W., et al. (2015). Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat. Commun.* 6, 8322. doi: 10.1038/ncomms9322
- Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810. doi: 10.1038/nature06244
- United States Pharmacopeial Convention (2015). Food Chemicals Codex, 9th Edn., 3rd Suppl. Rockville, MD: USP.
- Weese, J. S., and Martin, H. (2011). Assessment of commercial probiotic bacterial contents and label accuracy. *Can. Vet. J.* 52, 43–46.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

WM, AH, BZ, and BS are all employees of DuPont Nutrition & Health, which produces probiotic cultures that are used as ingredients in finished dietary supplements. Samples tested in this report were chosen without specific knowledge of probiotic supplier, and many products do not contain DuPont strains.

Copyright © 2016 Morovic, Hibberd, Zabel, Barrangou and Stahl. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Characterization of Gut Microbiome Dynamics in Developing Pekin Ducks and Impact of Management System

Aaron A. Best^{1*}, Amanda L. Porter¹, Susan M. Fraley^{1,2} and Gregory S. Fraley¹

¹ Department of Biology, Hope College, Holland, MI, USA, ² South Crossing Veterinary Center, Caledonia, MI, USA

OPEN ACCESS

Edited by:

Jennifer Ronholm,
Health Canada, Canada

Reviewed by:

Victor Satler Pyro,
University of São Paulo, Brazil
Kevin R. Theis,
Wayne State University School of
Medicine, USA

*Correspondence:

Aaron A. Best
best@hope.edu

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 12 August 2016

Accepted: 16 December 2016

Published: 04 January 2017

Citation:

Best AA, Porter AL, Fraley SM and
Fraley GS (2017) Characterization of
Gut Microbiome Dynamics in
Developing Pekin Ducks and Impact
of Management System.
Front. Microbiol. 7:2125.
doi: 10.3389/fmicb.2016.02125

Little to no research has been conducted on the gut microbiome of the Pekin duck, yet over 24.5 million ducks are raised for human consumption each year in the United States alone. Knowledge of the microbiome could lead to an understanding of the effects of growing conditions such as the use of prebiotics, probiotics, and enzymes in feeding practices, the use of antibiotics, and the sources of pathogenic bacteria in diseased ducks. In order to characterize changes in the caecal microbiome that occur as ducks develop through a typical industry grow-out period, a 16S rRNA community analysis of caecal contents collected over a 6-week period was conducted using a next generation sequencing approach. Transitions in the composition of the caecal microbiome occurred throughout the lifespan, with a large shift during days 4 through 10 posthatch. Two major phyla of bacteria were found to be present within the caeca of aviary raised ducks, with the relative abundance of each phylum varying by age of the duck. Proteobacteria is dominant for the first 3 days of age, and Firmicutes increases and dominates beginning at day 4. Barn raised ducks contained a significant population of Bacteroidetes in addition to Proteobacteria and Firmicutes at later developmental time points, though this phylum was absent in aviary raised ducks. Genera containing pathogens of anseriformes most often found in industry settings were either absent or found as normal parts of the caecal microbial populations. The high level differences in phylum abundance highlight the importance of well-designed sampling strategies for microbiome based studies. Results showed clear distinctions between Pekin Duck caecal contents and those of Broiler Chickens and Turkey in a qualitative comparison. These data provide a reference point for studies of the Pekin Duck through industry grow-out ages, provide a foundation for understanding the types of bacteria that promote health, and may lead to improved methods to increase yields and decrease instances of disease in agricultural production processes.

Keywords: pekin duck, duck microbiota, 16S rRNA amplicon sequencing, industry grow-out development, aviary environment, barn environment, bacteroidetes, riemerella

INTRODUCTION

Investigation of microbes associated with host organisms has become an increasingly important approach to better understand the host organisms, the microbial communities, and the interactions that occur between hosts, their microbes, and environment (Gilbert et al., 2016). Collectively, the microbes found in a particular environment can be referred to as a microbiome. The microbiome

of the gut of many organisms, in particular mammals, has been found to protect against pathogens, impact digestion, influence immune system function, and affect the health of individuals (e.g., Turnbaugh et al., 2006; Cho and Blaser, 2012; Flint et al., 2012; D'Argenio and Salvatore, 2015; Gilbert et al., 2016). The microbiome and its associated genetic content has been proposed to be an extension of the host organism that readily influences development and normal function and may be heritable (Ley et al., 2008; Funkhouser and Bordenstein, 2013; D'Argenio and Salvatore, 2015; Van Opstal and Bordenstein, 2015). Thus, the microbiome can be indicative of the health state of an individual, potentially linked to the absence or presence of disease, and suggest alterations in diet or treatment of gut related disease (Nicholson et al., 2005; Frank et al., 2007; Reid et al., 2011; Gevers et al., 2014; Lewis et al., 2015). A recent meta-analysis of microbiomes associated with various avian species revealed a dynamic, intestinal microbiome that changes with species of bird, host site of sample acquisition (e.g., crop, caecum), captivity status (wild or domesticated), and potential associations with diet (Waite and Taylor, 2014). Studies of commercial avian species have revealed changes in the microbiome associated with species, age, diet, host site, and commercial environmental conditions (van der Wielen et al., 2002; Lu et al., 2003; Gong et al., 2007; Wei et al., 2013; Choi et al., 2014; Stanley et al., 2014; Vasai et al., 2014; Roto et al., 2015). The majority of bacteria associated with avian species has been found in the intestinal caeca, where a relatively lower oxygen partial pressure and decreased enzyme and bile salt concentrations create conditions suitable for a variety of bacteria (Gabriel et al., 2006).

It is known that avian intestinal contents are much different than those of monogastric mammals (Pérez de Rozas, 2004), however most research has centered upon the gut microbiome of galliformes, specifically turkeys and broiler chickens. Broiler chicken caeca, analyzed using 16S rRNA clone libraries and Sanger sequencing, are dominated by Firmicutes at all ages. At days 7 through 14 of age, the chicken caecal contents resemble that of the chicken ileum. From day 14 forward, the compositions of the two regions diversify and stabilize, becoming significantly different from each other (Lu et al., 2003). By day 28, 5–10% of the caecal contents is composed of Bacteroidetes. Comparable results were found in 28-day old chickens using pyrosequencing to analyze the V1–V3 region of 16S rRNA (Choi et al., 2014; Stanley et al., 2014). In contrast, the caecal microbiome of 18-week old turkeys is dominated by Bacteroidetes (52%), while Firmicutes composition is 33% (Scupham et al., 2008). A recent study assessing 12–14 week old Pekin (*Anser platyrhynchos*) and Muscovy (*Cairina moschata*) ducks revealed that the caecal microbiomes consists of ~65 and ~50% Bacteroidetes, respectively, similar to the composition of turkeys at older ages (Vasai et al., 2014). Pathogens are often found in the gut of vulnerable galliformes. These include *Brachyspira*, causing colitis (Neo et al., 2013), *Campylobacter jejuni*, a common food-borne pathogen, and *Clostridium perfringens*, causing necrotic enteritis (Van Immerseel et al., 2004). Often present in broiler chickens and turkey, these pathogens have not been linked to anseriformes used in the food industry, such as the Pekin duck. *Riemerella anatipestifer* is the most common pathogen in anseriformes

(Wobeser, 1997), occurring globally in both commercially raised and wild ducks (Brogden, 1989). Other bacterial pathogens commonly associated with commercially raised ducks include *Escherichia coli*, *Salmonella*, *Streptococcus* and *Enterococcus*.

To date, very little research has been conducted on the gut microbiome of the Pekin duck, yet over 24.5 million ducks are raised for human consumption each year in the United States alone (AGMRC, 2012). Knowledge of the duck caecal microbiome could lead to a better understanding of the effects of management practices such as the use of prebiotics, probiotics and enzymes in feeding practices, the use of antibiotics, and to a better understanding of the sources of pathogenic bacteria in diseased ducks. This study characterized the microbiome of Pekin ducks over the industry standard 36-day period in which the ducks reach market weight, referred to as the grow-out period; determined if bacterial groups consistent with common anseriform pathogens are part of the endogenous caecal flora of developing ducks; and qualitatively compared the microbiomes of galliformes and anseriformes. Further, we compared the microbiome of ducks raised in a highly controlled aviary environment to a barn environment used in commercial practices to identify differences in microbiome composition related to environmental setting.

MATERIALS AND METHODS

Sample Collection

The study was conducted in two aviary experiments at Hope College in Holland, Michigan, and in a third barn experiment conducted at Maple Leaf Farms (MLF, Leesburg, IN, USA). A straight run (defined as a roughly equal mix of male and female) of day-old ducklings was obtained from MLF and housed in a controlled aviary setting at Hope College. The ducklings were of the commercial strain developed and utilized for international meat production by MLF. Housing conditions adhered to industry standards for 18:6 light:dark cycle, temperature (~18.5°C), humidity (60–65%), *ad libitum* access to commercial feed (identical feed provided by Maple Leaf Farms, Inc. for all studies) and pin-metered water lines, and pine litter flooring. Flock density was standardized across three pens based on industry standards (~0.16 m²/duck). All care and procedures were in concordance with the Guide for the Care and Use of Agricultural Animals in Research and Teaching (McGlone et al., 2010) and approved by the Hope College Animal Care and Use Committee (HCACUC).

Aviary Study 1. Sixty-two ducks were used. Ten ducks were sacrificed on days 1, 8, 15, 22, and 36, and 12 ducks on day 29. **Aviary Study 2.** Sixty-one ducks were used in which six ducks were sacrificed daily on days 1–10, excluding day 9, when 7 ducks were sacrificed. A final live weight was determined for each animal when euthanized. Samples of water, feed and bedding were also obtained for Aviary Studies 1 and 2. **Barn Study 3.** In an attempt to approximate the conditions of an actual commercial barn setting, our study was conducted in two research barns owned by Maple Leaf Farms, Inc. (Leesburg, IN USA). Each barn was divided into 4 equal sized pens with 1000 ducks per pen (~ 0.17 m² per duck). The study ran for the

duration of a typical grow-out period in the USA, approximately 36 days. After the first study was completed, the experiment was replicated thus providing a final $N = 16$ pens. The ducks used in the study were from the same commercial Pekin strain developed by Maple Leaf Farms, Inc. used for the aviary studies. Ducklings were randomly selected for both barns and placed within hours (hr) of hatch (day 1). After an initial 10-day brooding period in approximately one-third of the pen, they were given access to the entire floor space in each pen. Gut ecology samples were obtained on days 5, 23, and 33. In each pen, 3 apparently healthy ducks ($n = 12$ per barn) were selected at random and immediately euthanized using Fatal Plus (400 mg/kg pentobarbital, intraperitoneal). Pentobarbital is a well-known inhibitor of gastrointestinal motility. The ducks were weighed and the paired caeca of each animal were removed aseptically, and caecal contents were obtained and stored at -80°C until they were processed for microbial DNA analyses. Samples of water were also obtained for Barn Study 3. When the ducks reached targeted commercial weight (~ 3.5 kg) at 34 days, they were processed at the Maple Leaf Farms processing facility. The Hope College Animal Care and Use Committee approved all studies.

Bacterial DNA Isolation

Caecal contents (300 mg wet weight) and environmental samples (water, bedding, feed) were prepared for total community analysis using the PowerLyzer PowerSoil DNA Isolation Kit (MoBio, Carlsbad, CA, USA) according to the manufacturer's protocol except that MP Biomedical FastPrep24 lysing matrix D tubes with 1.4 mm ceramic spheres were used in place of the MoBio glass bead tubes for sample preparation, resulting in consistently higher DNA yields. DNA was eluted in a final volume of 100 μL of elution buffer according to the manufacturer's protocol. Total community DNA was stored at -20°C .

Sequencing of 16S rRNA

Total community DNA samples were submitted to the Institute for Genomics and Systems Biology Next Generation Sequencing (IGSB-NGS) Core Facility at Argonne National Laboratory for sequencing of community 16S rRNA genes. Briefly, genomic DNA was amplified using the Earth Microbiome Project barcoded primer set, adapted for the Illumina MiSeq (Caporaso et al., 2012). The V4 region of the 16S rRNA gene (515F-806R) was amplified with region-specific primers that included the Illumina flowcell adapter sequences and unique, 12 base barcode sequences. Each 25 μl PCR reaction contained 12 μl of MoBio PCR Water (Certified DNA-Free), 10 μl of 5 Prime HotMasterMix (1x), 1 μl of Forward Primer (5 μM concentration, 200 pM final), 1 μl Golay Barcode Tagged Reverse Primer (5 μM concentration, 200 pM final), and 1 μl of template DNA. The conditions for PCR were as follows: 94°C for 3 min to denature the DNA, with 35 cycles at 94°C for 45 s; 50°C for 60 s; and 72°C for 90 s, with a final extension of 10 min at 72°C to ensure complete amplification. The PCR amplifications were done in triplicate, and then pooled. Following pooling, amplicons were quantified using PicoGreen (Invitrogen) and a plate reader. Once quantified, different volumes of each of the products were pooled into a single tube so that each amplicon is represented equally.

This pool was then cleaned up using an UltraClean® PCR Clean-Up Kit (MoBIO), and then quantified using the Qubit (Invitrogen). After quantification, the molarity of the pool was determined and diluted down to 2 nM, denatured, and then diluted to a final concentration of 6.75 pM with a 10% PhiX spike for sequencing on the Illumina MiSeq.

Data Analysis

The Quantitative Insights into Microbial Ecology (QIIME) software package, version 1.9.1 (Caporaso et al., 2010b) was used to analyze 16S microbial sequencing data. We utilized custom shell scripts to perform "upstream" and "downstream" processing stages as recently described (Navas-Molina et al., 2013). All steps requiring comparison of sequences to a reference database used the GreenGenes database, release 13_8 (DeSantis et al., 2006). For the upstream analysis steps, we performed demultiplexing and quality-filtering for Illumina based sequence reads using default values. Clustering of sequencing reads that passed quality filters into operational taxonomic units (OTUs) was performed through an open-reference strategy at a threshold of 97% identity, using uclust (Edgar, 2010). Taxonomic assignment of representative OTUs was performed using the QIIME rtax workflow in order to take advantage of paired end sequencing reads (Soergel et al., 2012). The rtax settings allowed for inclusion of OTUs identified by non-paired reads (-single_ok option). Chimeric sequences were removed using ChimeraSlayer (Haas et al., 2011). In order to construct a phylogenetic tree of the identified OTUs, sequences were aligned using PyNAST (Caporaso et al., 2010a) against the GreenGenes core set template (DeSantis et al., 2006; McDonald et al., 2012). A phylogenetic tree was constructed using FastTree 2 (Price et al., 2010) within the QIIME workflow. Finally, an OTU table in BIOM format (McDonald et al., 2012) was produced, along with a complete metadata mapping file for use in downstream analysis steps.

Alpha diversity metrics (observed species, phylogenetic distance, Good's coverage, Chao1, and Shannon) were performed on all samples at the maximum depth for each sample to yield summary statistics for the data set using QIIME and are reported in Supplementary Table 1. We performed secondary filtering of OTUs to minimize the effect of very low abundance OTUs, using the recommended value of <0.005% of the total number of sequences (Bokulich et al., 2013) as a conservative threshold for removal of OTUs from further consideration. The filtered BIOM table, phylogenetic tree and metadata sample table were passed to the core diversity analysis workflow (Lozupone and Knight, 2005; Navas-Molina et al., 2013) in QIIME to perform taxa summarization, alpha diversity, beta diversity, and taxon differential distribution analyses. Full output files from taxa summary analyses performed in QIIME are reported in Supplementary Files 1–5. A jackknifed beta diversity analysis (Lozupone et al., 2011) was conducted to assess statistical variation of sample location in principal coordinate analysis (PCoA) plots based on unweighted and weighted UniFrac distances. We used EMPeror (Vázquez-Baeza et al., 2013) to visualize PCoA plots. Following initial evaluation of the data, the BIOM table was variously filtered to focus on particular sample comparisons as described in the Results section; rarefaction

for all analyses was to 10,000 reads per sample. Targeted group significance tests were used to compare OTU frequencies amongst combinations of age, gender, and experimental setting as implemented in: QIIME (Kruskal-Wallis, Mann-Whitney U, group and pairwise adonis); IBM SPSS version 23 for Macintosh (repeated measures ANOVA); and the R packages phyloseq (McMurdie and Holmes, 2013) and vegan (Dixon, 2003) (DESeq2, two-way adonis). Sequencing data have been deposited as a combined data set for all three studies in the European Nucleotide Archive (ENA) under the accession number (<http://www.ebi.ac.uk/ena/data/view/PRJEB15658>).

RESULTS

Microbial communities associated with caecal contents of Pekin ducks were analyzed in order to understand the structure of and changes associated with the development of ducks through a typical industry growth period. Two studies were conducted in a controlled aviary setting—Aviary Study 1 was designed as a broad overview of the 36-day developmental cycle; Aviary Study 2 was designed to focus on the first 10 days of duck development. These data were compared to Barn Study 3, a parallel study that occurred in a production barn environment (Schenk et al., 2016) in order to examine differences in microbial community structure and developmental changes that could occur in different environments. The microbial community profiles were searched for the presence and abundance of potential anseriform pathogens. Processing of sequencing data and evaluation of taxonomic distribution, alpha diversity, and beta diversity metrics were performed in QIIME. Basic information for samples, including barcodes used, sequencing reads, number of taxa observed, estimates of alpha diversity, and metadata are reported in Supplementary Table 1. Clear shifts in the microbial communities occurred in all three studies associated with the age of the ducks and the environmental setting, whereas significant associations with other factors, such as the sex of the ducks, were not observed.

Aviary Study 1: Differences in Microbial Populations Associated with Age of the Developing Duck

In order to assess the microbial populations of ducks throughout a typical industry grow-out period, we collected caecal contents at 7-day intervals through 36 days for analysis via 16S rRNA gene sequencing of total community DNA. Multiple alpha diversity metrics revealed clear differences among microbial populations in ducks of different ages through the grow-out period with respect to richness and diversity. In general, there is a significant increase in the diversity of the microbial populations by all metrics as ducks mature (Supplementary Figure 1A, repeated measures ANOVA inset for each metric). Ducks at day 15 or less have fewer than 80 observed species, whereas ducks at day 22 or older have greater than 115 observed species. Ducks at day 36 had, on average, 143 observed species. All pairwise comparisons of the number of observed species grouped by day were statistically significant (Supplementary Table 2, pairwise

t-tests, $p < 0.05$) with the exception of day 1 vs. day 15 and day 22 vs. day 29. In fact, all alpha diversity metrics produced statistically indistinguishable values for day 22 vs. day 29. When considering the Shannon diversity metric, which takes into account richness and evenness of species, a pattern of early and late stages in the 36 day grow-out period emerges. This is supported by statistically significant differences between early (Days 1 and 8) and late (Days 15, 22, 29, and 36) age duck samples (Supplementary Table 2, pairwise *t*-tests, $p < 0.05$), and the absence of statistically significant differences between days within the early and late groupings (Supplementary Table 2, pairwise *t*-tests, $p > 0.05$). This suggests that there are early and late stages in the 36-day grow-out period that are distinct from each other.

The structure of the duck caecal microbial populations are distinct and are clearly correlated with the age of the duck based on beta diversity measures. In weighted UniFrac Principal Coordinates Analysis (PCoA), the first principal coordinate (PC1) explained 41% of the variation among samples, with PC2 and PC3 explaining 28 and 10% of the variation, respectively. This analysis shows distinct clustering of individual duck caecal samples associated with age (Figure 1), supports the distinction between early and late age ducks, and suggests a difference between Day 1 and Day 8 ducks. Differences among age groups were shown to be statistically significant in a multivariate ANOVA based on dissimilarities (adonis) test (weighted UniFrac distances, $DF = 5, 999$ simulations, $F = 41.904$, $R^2 = 0.79$, $p = 0.001$) All pairwise comparisons of age groups were statistically significant (pairwise adonis, weighted UniFrac distances, $DF = 1, 999$ simulations, $F = 5.48\text{--}195.24$, $R^2 = 0.24\text{--}0.91$, $p = 0.001$).

Aviary Study 1: Major Shift in Microbial Caecal Contents Occurs Early in the 36 Day Grow-Out Period

Taxa summaries of the samples grouped by age from Aviary Study 1 revealed a shift in phylum level relative abundances by the Day 8 sampling point (Figure 2A). Day 1 ducks were dominated by the phylum Proteobacteria, ranging from 77 to 99% of the microbial population in an individual. By day 8, the population had shifted to dominance by the phylum Firmicutes, ranging from 81 to 98% of the population in an individual. The dominance of Firmicutes extended through the rest of the grow-out period, making up an average of 96% of the microbial population.

The shift to Firmicutes by day 8 and subsequent maintenance of this shift suggested that a major transitional period in caecal population development occurred prior to day 8. However, it was also apparent from the data that day 8 individuals were different in both the diversity of taxa present and the composition of the population with respect to days 15 through 36. This is borne out by examining the taxonomic composition at ranks below the phylum level, where it is clear that day 8 individuals had distinct relative abundances of different taxa (Figure 2B). In order to further characterize the differences in early and late age ducks, taxonomic groups that are significantly differentially distributed between age groups were determined using DESeq2 (Anders and Huber, 2010) as implemented in phyloseq (McMurdie and Holmes, 2013) (Supplementary Table 3). In pairwise comparisons

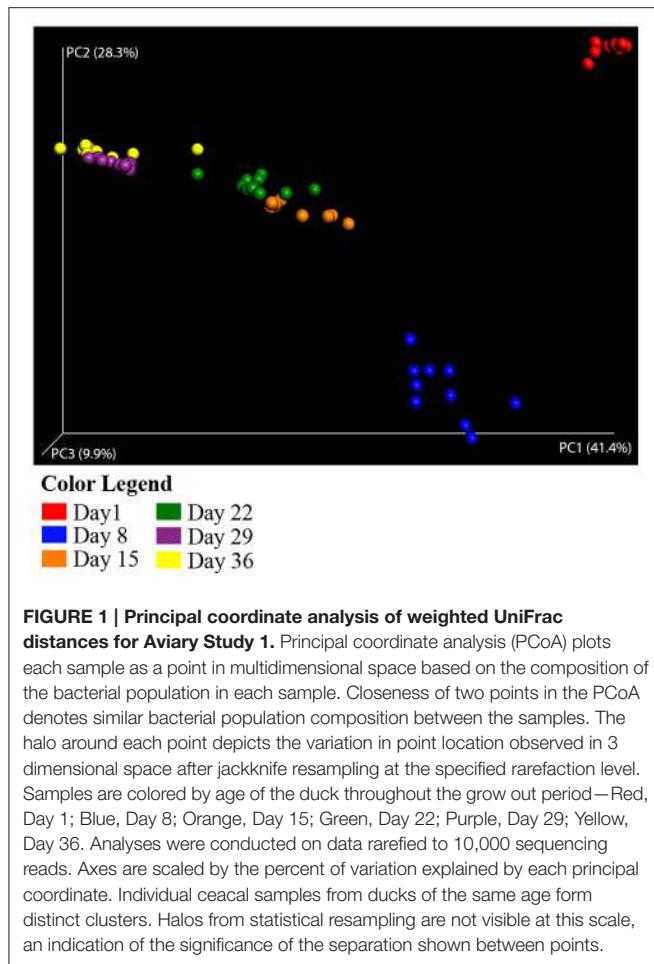


FIGURE 1 | Principal coordinate analysis of weighted UniFrac distances for Aviary Study 1. Principal coordinate analysis (PCoA) plots each sample as a point in multidimensional space based on the composition of the bacterial population in each sample. Closeness of two points in the PCoA denotes similar bacterial population composition between the samples. The halo around each point depicts the variation in point location observed in 3 dimensional space after jackknife resampling at the specified rarefaction level. Samples are colored by age of the duck throughout the grow out period—Red, Day 1; Blue, Day 8; Orange, Day 15; Green, Day 22; Purple, Day 29; Yellow, Day 36. Analyses were conducted on data rarefied to 10,000 sequencing reads. Axes are scaled by the percent of variation explained by each principal coordinate. Individual caecal samples from ducks of the same age form distinct clusters. Halos from statistical resampling are not visible at this scale, an indication of the significance of the separation shown between points.

with late age ducks (days 15, 22, 29, and 36), an average of 43 OTUs were identified as significantly enriched (Benjamini-Hochberg adjusted $p \leq 0.05$) in day 1 ducks. These OTUs were evenly distributed between the Proteobacteria and Firmicutes in the comparison with day 15, but the distribution was increasingly biased toward members of the Proteobacteria in comparisons with days 22, 29, and 36. The OTUs identified as enriched in the late age ducks as compared to day 1 ducks were almost all members of Firmicutes, with a small number distributed between Actinobacteria and Tenericutes. The largest numbers of enriched OTUs identified in comparisons with day 1 ducks were for days 29 (105 OTUs) and 36 (104 OTUs). In pairwise comparisons of late age ducks and day 8 ducks, an average of 56 OTUs were identified as significantly enriched in day 8 ducks. These OTUs were roughly evenly distributed between Proteobacteria and Firmicutes for all late age comparisons, with a trend toward Firmicutes in comparisons with days 29 and 36. As with the day 1 duck comparisons, the OTUs identified as enriched in the late age ducks as compared to day 8 ducks were almost all Firmicutes, with a small number distributed between Actinobacteria and Tenericutes. The largest number of enriched OTUs in comparisons with day 8 ducks were for days 29 (97 OTUs) and 36 (95 OTUs). A pairwise comparison

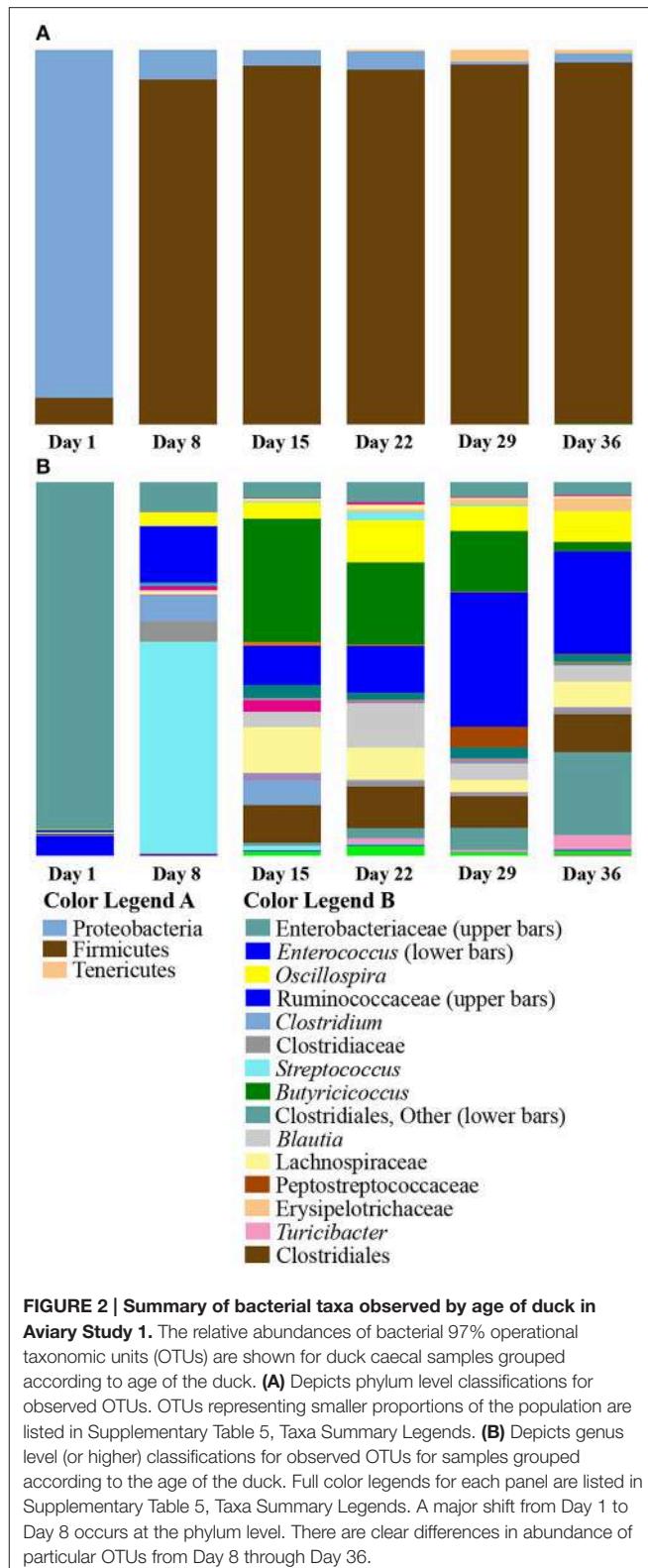


FIGURE 2 | Summary of bacterial taxa observed by age of duck in Aviary Study 1. The relative abundances of bacterial 97% operational taxonomic units (OTUs) are shown for duck caecal samples grouped according to age of the duck. **(A)** Depicts phylum level classifications for observed OTUs. OTUs representing smaller proportions of the population are listed in Supplementary Table 5, Taxa Summary Legends. **(B)** Depicts genus level (or higher) classifications for observed OTUs for samples grouped according to the age of the duck. Full color legends for each panel are listed in Supplementary Table 5, Taxa Summary Legends. A major shift from Day 1 to Day 8 occurs at the phylum level. There are clear differences in abundance of particular OTUs from Day 8 through Day 36.

of day 1 and day 8 ducks identified 36 OTUs and 29 OTUs as significantly enriched for each day, respectively. The day 1 enriched OTUs were evenly distributed between Proteobacteria

(17 OTUs) and Firmicutes (19 OTUs), whereas the day 8 enriched OTUs were predominantly from Firmicutes (26 OTUs). For both days, all enriched Proteobacteria OTUs were members of the Enterobacteriaceae; all but one of the enriched Firmicutes OTUs were members of either Bacilli or Clostridiales classes. These results are consistent with the shift to Firmicutes early in the grow-out period and with increased alpha diversity metrics of late aged ducks.

Aviary Study 2: Analysis of First 10 Days of Developmental Grow-Out Period

The data from Aviary Study 1 led us to focus on the first 10 days of the grow-out period in a second group of ducks (Aviary Study 2) in order to more fully characterize the shift from Proteobacteria to Firmicutes. Alpha diversity metrics showed a consistent level of diversity throughout the 10 day period; three of the metrics showed statistically significant differences among age groups (Supplementary Figure 1B, repeated measures ANOVA inset for each metric). The Shannon diversity index exhibited the strongest pattern and showed statistically significant differences between pairwise combinations of early (Days 1, 2, and 3) and late (Days 5, 6, 8, 9, and 10) portions of the study period (*post-hoc t*-tests, Bonferroni corrected $p < 0.05$). The number of observed species ranged from 67 to 93 through the 10 days, consistent with that seen in the early days from Aviary Study 1.

The structures of the duck caecal populations were not as clearly distinguished in PCoA plots, in contrast to the broader age range covered in Aviary Study 1. The trend observed in PC1 (49% of the data explained) appeared to loosely correlate with age (Supplementary Figure 2), and the groupings by age were statistically supported in both weighted (adonis, $DF = 9$, 999 permutations, $F = 7.7439$, $R^2 = 0.58$, $p = 0.001$) and unweighted UniFrac analyses (adonis, $DF = 9$, 999 permutations, $F = 1.9546$, $R^2 = 0.26$, $p = 0.001$).

Despite weaker trends in PCoA analyses, the summaries of the taxonomic groups observed in age-based categories of ducks recapitulated the major shift from Proteobacteria to Firmicutes by day 8 in the Aviary Study 1. Proteobacterial dominance persisted through the first 2 days of age (Figure 3A; this pattern held for all individuals from days 1 and 2 in the study, Figure 3B). The phylum Firmicutes rose in abundance sharply between days 2 and 3 of age, with an increase from averages of 11 to 48% of the population. Day 3 represented a clear transition—3 out of 6 individuals maintained proteobacterial dominance, and 3 individuals had already shifted to dominance by Firmicutes (Figure 3B). By day 4, the proportion of Firmicutes rose to an average of 66% and stabilized from day 5 through day 10 of the second study at ~78% of the population. Days 2 and 4 were shown to be significantly different in pairwise comparisons of weighted UniFrac distances (adonis, $DF = 1$, 999 permutations, $F = 18.637$, $R^2 = 0.65$, $p = 0.005$) along with 31 of 45 possible pairwise combinations between Days 1–10 (adonis, $DF = 1$, 999 permutations, $F = 2.7867$ – 32.345 , $R^2 = 0.22$ – 0.76 , $p < 0.05$).

The transition from Proteobacteria to Firmicutes dominance from days 3 through 10 was characterized by the increase of a small number of major taxonomic groups, including the classes

Bacilli, Clostridia, and Erysipelotrichi. The Clostridia comprised the majority of the Firmicutes observed in most individuals from days 3 through 10, ranging from 45 to 78% of the population (Figure 3C). Comparatively smaller populations of Bacilli and Erysipelotrichi existed through this period, though they both rose to over 10% of the population in some individuals.

The dominant taxa within the class Clostridia were members of the families Lachnospiraceae, genus *Blautia* (Clostridiales, Lachnospiraceae); Clostridiaceae, genus *Clostridium* (Clostridiales, Clostridiaceae), and an uncharacterized genus in the family Clostridiaceae; and Ruminococcaceae, genera *Oscillospira* (Clostridiales, Ruminococcaceae) and *Butyrivibrio* (Clostridiales, Ruminococcaceae) (Figure 3D). The genus *Blautia* comprised less than an average of 1% of the population through day 5, but jumped to a peak of 25% of the population on day 6. Through day 10, the genus ranged from 13 to 25% of the population. The genus *Clostridium* was present at less than 1% of the population through day 3, followed by a rise to 13% on day 4, a peak of ~30% on days 6 and 7, and stabilizing at ~17% of the population through day 10. In contrast, the uncharacterized Clostridiaceae genus was present as a large fraction, ~5%, of the population of ducks from day 1, expanded to be ~42% of the population on days 3–5 of the grow-out period, and declined to ~10% of the population on days 6–8 and 10, ranging down to 3% on day 9 (Figure 3D). Thus, in days 3–5, the large increase in Firmicutes is due primarily to a single genus in the Clostridiaceae. The genus *Oscillospira* is less than 1% of the population through day 6, but ranges from 4 to 8% of the population on days 7–10, ending at 5% by day 10. *Oscillospira* represents the only major taxon of the family Ruminococcaceae until day 10, when the genus *Butyrivibrio* blooms to become 9% of the population. Prior to day 10, *Butyrivibrio* is present at less than 0.5% of the population. Each of these groups is often associated with the gut microbiome from a variety of animals (Biddle et al., 2013; Tims et al., 2013; Eren et al., 2014; Geirnaert et al., 2015) and serve as examples of the individual taxon dynamics that occur during the development of ducks.

Overview of the Full 36 Day Grow-Out Period

The combination of Aviary Studies 1 and 2 reveal a clear succession of microbial populations through a highly variable early stage to a more stable late stage of development. The fine grained sampling through the first 10 days of development shows that the transition from dominance by Proteobacteria to dominance by Firmicutes occurs by day 4, however there are clear differences in the types of Firmicutes and their relative abundances observed as ducks mature.

Major taxa seen in the first 10 days of the grow-out period also appear as part of the populations seen in late age ducks from Aviary Study 1. Following these major taxa through the rest of the grow-out period shows the marked shift from Proteobacteria to Firmicutes. The proteobacterial population that remains after the shift is dominated by an undefined genus comprising the same two 97% OTUs in both aviary studies. Both OTUs are significantly differentially distributed across ages

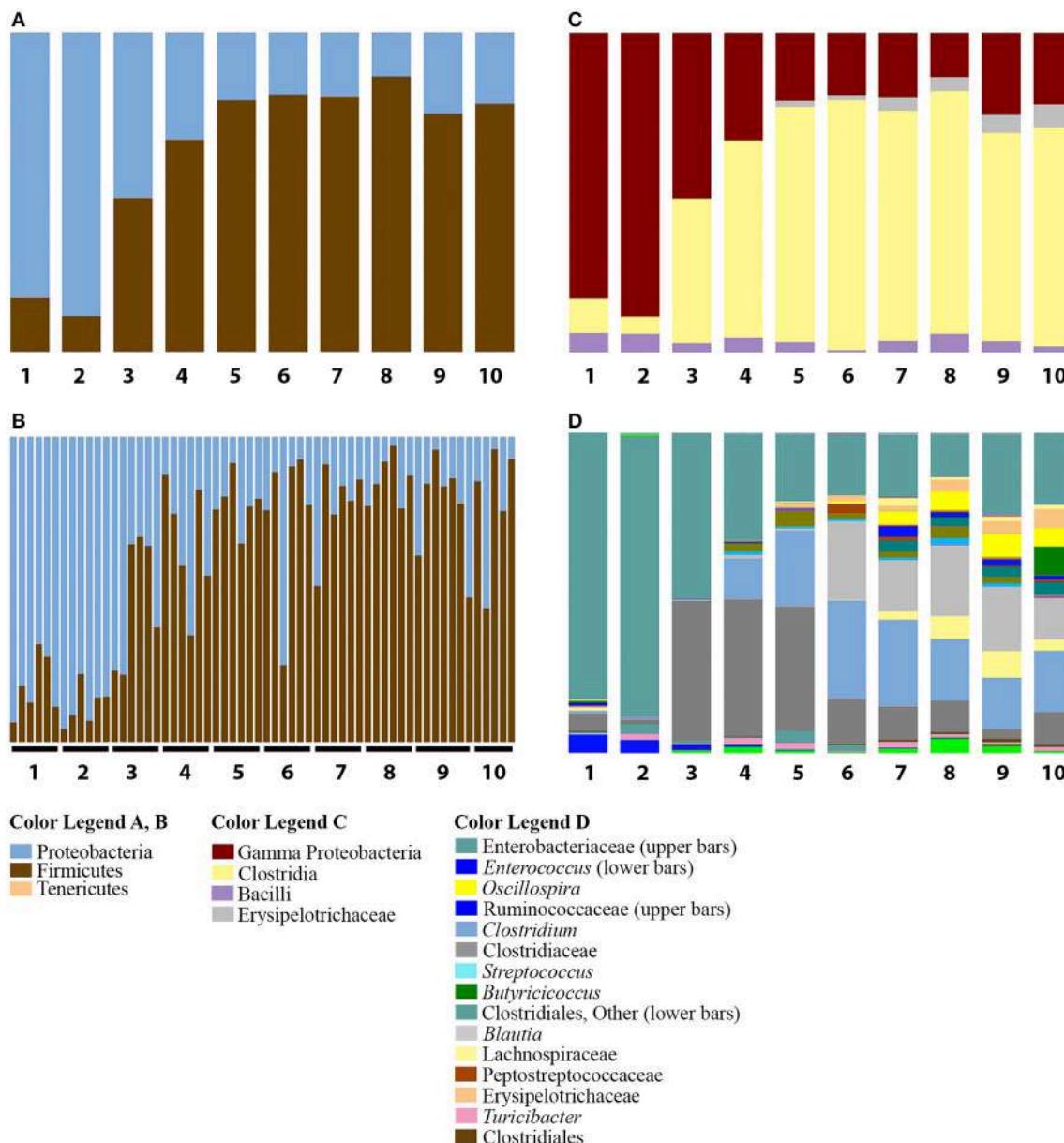


FIGURE 3 | Summary of bacterial taxa observed by age of duck in Aviary Study 2. The relative abundances of bacterial 97% operational taxonomic units (OTUs) are shown for duck caecal samples from Aviary Study 2, representing the first 10 days of the grow out period. **(A)** Depicts phylum level classifications for observed OTUs for samples grouped according to age of the duck. Numbers under each bar indicate the age of the ducks in days. **(B)** Depicts phylum level classifications for observed OTUs of individual samples, sorted according to age of the duck. Each horizontal line underneath the bars encompasses the samples derived from ducks of the indicated age in days. **(C)** Depicts class level classifications for observed OTUs for samples grouped according to age of the duck. **(D)** Depicts genus level (or higher) classifications for observed OTUs for samples grouped according to the age of the duck. Full color legends for each panel are listed in Supplementary Table 5, Taxa Summary Legends. The observed transition from Proteobacteria to Firmicutes is seen to occur by Day 3 and occurs in all individuals that make up the age groups **(A,B)**. The distribution of taxa below the phylum level shows that ages 4 and 5 are distinct from ages 6 to 10 **(C,D)**.

(Kruskal-Wallis, Bonferroni corrected $p = 7 \times 10^{-12}$). Within the Firmicutes, the genus *Blautia* peaks at 12% of the population in 22 day old ducks followed by stabilization of the population at ~4% by day 29 (Figure 2B). The percentages of the total population of *Blautia* between days 6 and 22 are well in excess of those seen in other organisms (Eren et al., 2014), though

this comparison is to developed, rather than to developing specimens. The family Ruminococcaceae is represented primarily by *Oscillospira* prior to day 10 in Aviary Study 2 ducks (Figure 3D) and is seen to be present in high proportions of the populations in Aviary Study 1 ducks from day 15 to the end of the grow-out period (Figure 2B). In contrast, *Butyrivibrio*

appears at day 10 in study 2 ducks, peaks at 33% of the population on day 15 of study 1 ducks, and remains above 15% of the population through day 29 before rapidly decreasing to 5% of the population by day 36 of the grow-out period (**Figures 2B, 3D**). Another group in the Ruminococcaceae family rapidly rises as a percent of the population in late age ducks, bringing the relative abundance of this family to a range of 19 to 60% from day 15 on (**Figure 2B**). The two major taxonomic groups within the family Clostridiaceae, *Clostridium* and an undefined genus, actually peak within the first 10 days at up to 19% and 10% of the population followed by a rapid decrease to become ~1% of the population or less by day 36 (**Figures 2B, 3D**). Thus, even though the phylum level Firmicutes population rises as age increases, the dominant taxa representing Firmicutes early in development are almost fully replaced by other Firmicutes late in development.

Both aviary studies were conducted 6 months apart and overlap with two time points (Day 1 and Day 8 ducks). A DESeq2 analysis of the microbial populations observed in both studies showed that there are 24 OTUs identified as significantly different between Day 1 ducks from both studies and 58 OTUs identified as significantly different between Day 8 ducks (Supplementary Table 3). The larger number of enriched OTUs identified for the day 8 comparison is consistent with the PCoA (Supplementary Figure 3); day 1 ducks from both studies cluster closely in PCoA (Supplementary Figure 3A), whereas day 8 ducks are clearly distinct between the two studies (Supplementary Figure 3). Despite the differences at day 8, many of the taxa observed in the 10 day period of study 2 were observed in later ages of study 1 ducks (cross reference Supplementary Figure 3B with results above).

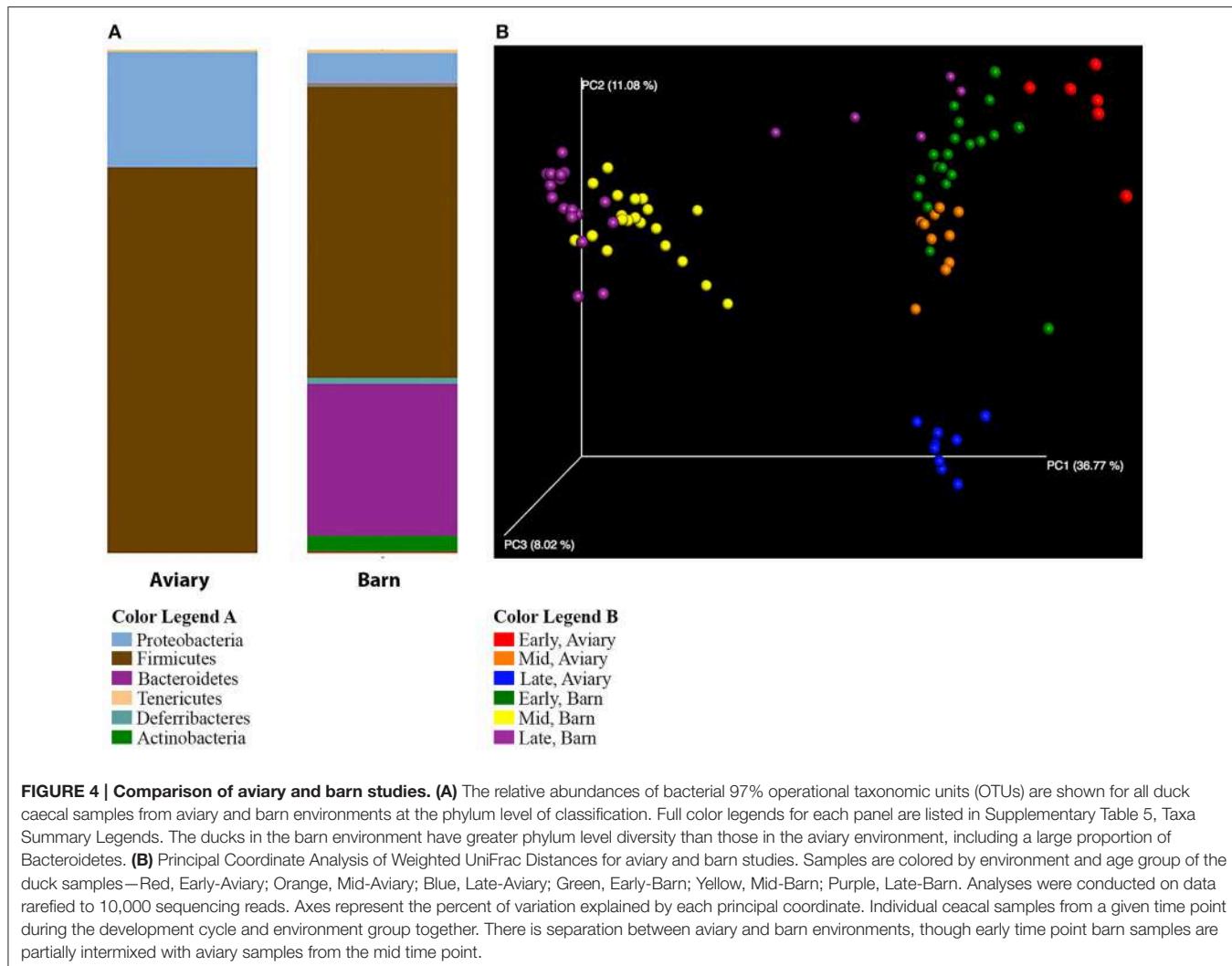
Barn Study 3: Bacteroidetes Absent from Duck Caecal Microbiome in Aviary Setting

Four major bacterial phyla most often associated with animal gut systems are Bacteroidetes, Firmicutes, Proteobacteria, and Actinobacteria (Ley et al., 2008). However, Bacteroidetes was not observed in the aviary data sets. To address whether this absence is due to the environmental setting or to being a unique feature of anseriformes, we analyzed data from a parallel study conducted in an agricultural barn setting (Schenk et al., 2016). The original goal of the parallel barn study (herein referred to as Barn Study 3) was to assess the advantages and disadvantages of watering the ducks with an open water trough or with a pin-metered line system. As such, a limited sampling of duck caecal contents was built into the design, restricted to days 5, 21, and 33 of the grow-out period. The source of the ducks in Aviary Studies 1 and 2 was the same source as the ducks in Barn Study 3. In the comparison of the ducks from the 3 studies, we used ducks from Aviary Study 2, day 5; Aviary Study 1, day 22; and Aviary Study 1, day 36. These ages were then grouped into early, mid and late time points in the grow-out period, respectively, and combined with the corresponding early, mid and late time points from the Barn Study 3 grow-out period for comparative analysis of the microbial populations.

Several observations confirm the expectation that the barn environment is very different from the aviary environment,

and are reflected in significant differences in the composition of the caecal microbiome of ducks in both settings. Alpha diversity measures show a marked increase in the diversity of the microbiome in barn-raised ducks compared to aviary-raised ducks. The Shannon entropy for aviary ducks was 3.1 and for barn ducks was 4.6, a statistically significant difference (*t*-test, $t = -8.12, p = 0.001$). The number of observed species is also significantly higher in barn-raised ducks (90 vs. 155, respectively; *t*-test, $t = -12.13, p = 0.001$). The distributions of taxa observed in aviary vs. barn-raised ducks are consistent with alpha diversity metrics (**Figure 4A**). As noted before, the major phyla in the aviary ducks are the Proteobacteria (24%) and the Firmicutes (76%). In contrast, the barn-raised ducks contain four major phyla found most often in other animal systems, Firmicutes (57%), Bacteroidetes (30%), Proteobacteria (6%), and Actinobacteria (3%). The differences in the barn and aviary duck caecal microbiomes are clearly seen in PCoA plots as clusters associated with both environmental setting and age (**Figure 4B**). The factors environment and age and the interaction between the two factors were shown to be statistically significant between the barn and aviary microbiomes (two-way adonis, Environment— $DF = 1, F = 46.961, R^2 = 0.25293, p = 0.001$; Age Group— $DF = 2, F = 18.294, R^2 = 0.19706, p = 0.001$; Environment: Age Group— $DF = 2, F = 6.061, R^2 = 0.06528, p = 0.001$). A DESeq2 analysis of the aviary and barn environments identified 217 OTUs as significantly different between the two environments, 77 significant to the Aviary and 140 significant to the Barn environment (Supplementary Table 4). Ninety-five percent of the OTUs identified in the barn environment are associated with the phyla Firmicutes, Bacteroidetes and Proteobacteria, whereas only Firmicutes and Proteobacteria are represented among enriched OTUs in the aviary environment. All but 8 of these OTUs are associated with the phylum Firmicutes. These data are consistent with increased alpha diversity metrics in the barn environment associated with major bacterial groups.

Subdividing the aviary and barn raised ducks into early, mid and late age groups reveals that the development progression observed in the aviary setting also takes place in the barn setting (**Figure 5**). In both early time points (day 5), Firmicutes is the dominant phylum (aviary, 79%; barn, 89%). This is consistent with the shift from Proteobacteria dominance in days 1 and 2 to Firmicutes dominance by day 4 in the aviary setting. The mid and late time points show significant increases in the population of Bacteroidetes in the barn setting, from 0.2% on day 5 to ~40% on days 21 and 33. Within the aviary setting, the numbers of significantly enriched OTUs in aviary-early and aviary-mid/aviary-late pairwise comparisons using DESeq2 are 126 and 116, respectively. For the same pairwise comparisons within the barn setting, the numbers of enriched OTUs are 223 and 219. Pairwise comparisons between the two environments matched by age grouping show 121 OTUs (aviary/barn early), 200 OTUs (aviary/barn mid), and 200 OTUs (aviary/barn late) as significantly enriched, with 77, 64, and 59% of OTUs being enriched in the barn environment in each comparison, respectively (Supplementary Table 4). In all age based comparisons between the environments, Bacteroidetes OTUs are identified as significantly enriched (early,

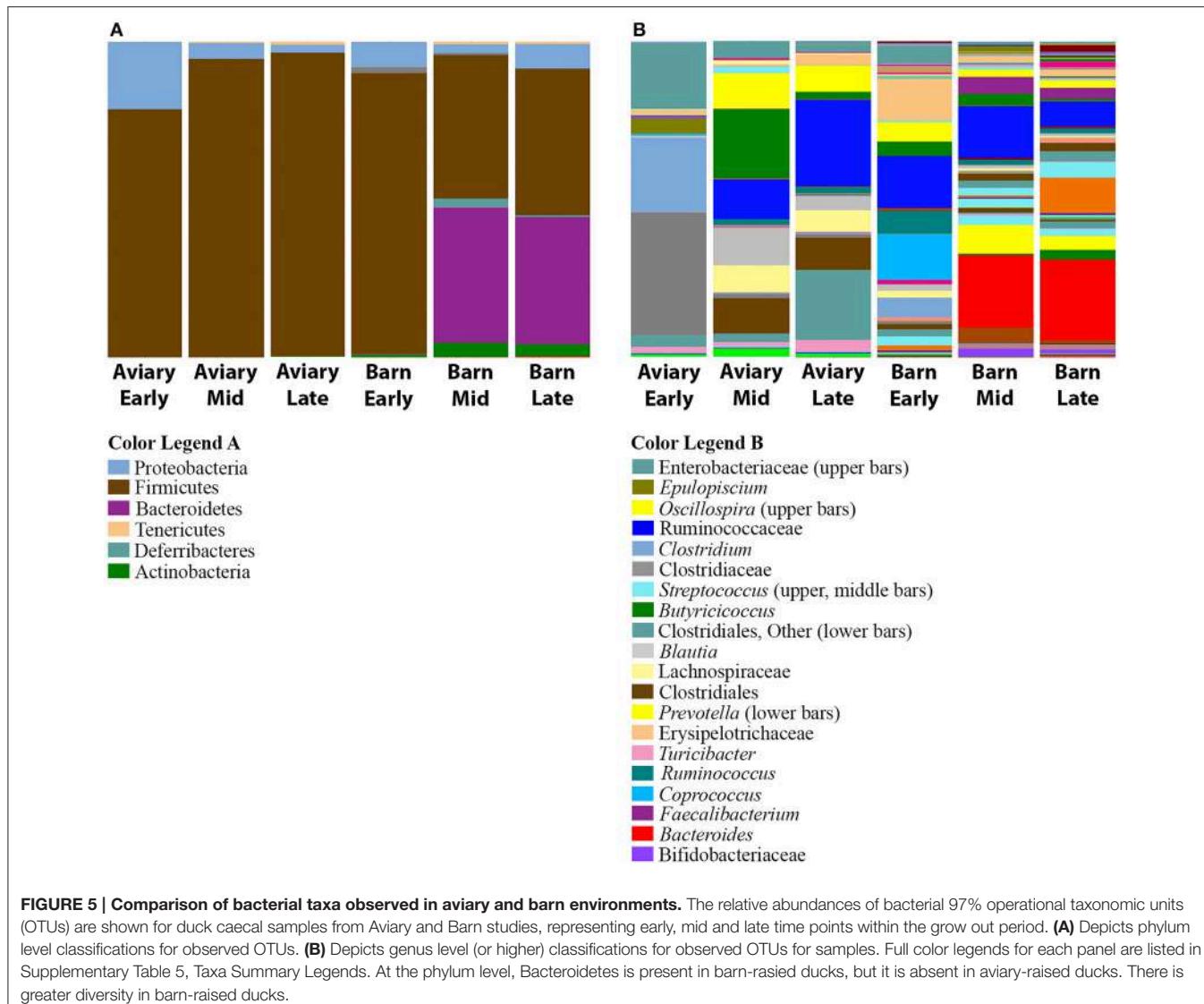


3; mid, 27; late, 29). The Firmicutes represents the largest fraction of significantly enriched OTUs in the barn setting at all age comparisons (early, 89%; mid, 59%; late, 49%), highlighting the variability at the OTU level seen between the two environments for major taxonomic groups. Despite the taxon level diversity between the environments, similar shifts in high level taxonomic ranks are observed in both environments. All pairwise comparisons within and between environments grouped by age are provided in Supplementary Table 4.

Evaluation of Common Anseriforme Bacterial Pathogens

The five most common bacterial pathogens associated with ducks in the production environment are *Riemerella anatipestifer*, *Escherichia coli*, *Salmonella*, *Streptococcus*, and *Enterococcus*. Of these, the focus of most monitoring efforts are *R. anatipestifer*, *E. coli*, and *Streptococcus* (personal communication with Dr. Dan Shafer, Vice-President Live Production, MLF, Inc.). The OTU assignments for ducks raised in the aviary and barn settings

were queried for these organisms. None of the assigned OTUs were identified as *R. anatipestifer*, *E. coli*, or *Salmonella*, however, assignments to the genus level for *Escherichia* are problematic with the greenegenes database used in this study and in many current microbiome studies (Nelson et al., 2014). Assignments to Enterobacteriaceae were present, which is the group harboring *E. coli* and *Salmonella*. Both *Streptococcus* and *Enterococcus* were identified as part of the microbiome of the duck caecum in both aviary and barn settings, and *Streptococcus* was found to be a dominant part of the population in some ducks. Taxa summaries of the ducks grouped by study and day show that the microbiomes of Aviary Study 1 ducks at Day 8 were dominated by a population of the genus *Streptococcus* (Firmicutes, Bacilli) making up 57% of the population (Supplementary Figure 3B). The *Streptococcus* population was found in Day 8 ducks from Aviary Study 2, but at a very low percentage of the total population ($9.2 \times 10^{-5}\%$). The *Streptococcus* genus that dominates Day 8 ducks from Aviary Study 1 is comprised of seven 97% OTUs, but a single OTU represents 99.8% of the population. This OTU is also present in ducks from Aviary Study 2, albeit



at a very small proportion of the total population in study 2 ducks, and it is significantly differentially distributed between the two studies (Mann-Whitney U, Bonferroni corrected $p = 0.002$) and across ages through the entire grow-out period (Kruskal-Wallis, Bonferroni corrected $p = 0.0004$). The same OTU is also identified in ducks from Barn Study 3 at low percentages of the population. Despite the presence of these taxa, all ducks in the three, independent studies were healthy and reached market weights by Day 36 of the grow-out period.

DISCUSSION

Commercial farming of Pekin ducks is a multibillion dollar industry worldwide, with the production of over 24 million ducks per year in the United States, alone. This study provides the first assessment of the microbial populations present in the caecum of a commercial strain of Pekin duck throughout a 36-day grow-out period that is typical of industry practices. In

particular, the developmental progression is characterized by a major transition from Proteobacteria in the first 2 days of age to dominance by Firmicutes by 5 days of age. This transition is observed in both aviary and industry barn settings, but there are stark differences in the diversity and composition of the microbial populations in the two settings. As discussed below, there are clear differences between anseriformes and galliformes, indicating that it is not advisable to extrapolate results of studies affecting the microbiomes from one bird order to another. These results highlight the necessity of careful experimental design as industries consider improvements to management practices based on nutrition, feed, prebiotic, probiotic, and antibiotic usage.

The caecae were utilized in this study because of their role in digestion and the overall health of birds. With internal villi, a sphincter and a blind end, nutrient-rich liquid contents of the digestive tract are concentrated in the caecae via reverse-peristalsis from the small intestine and colon (for reviews of

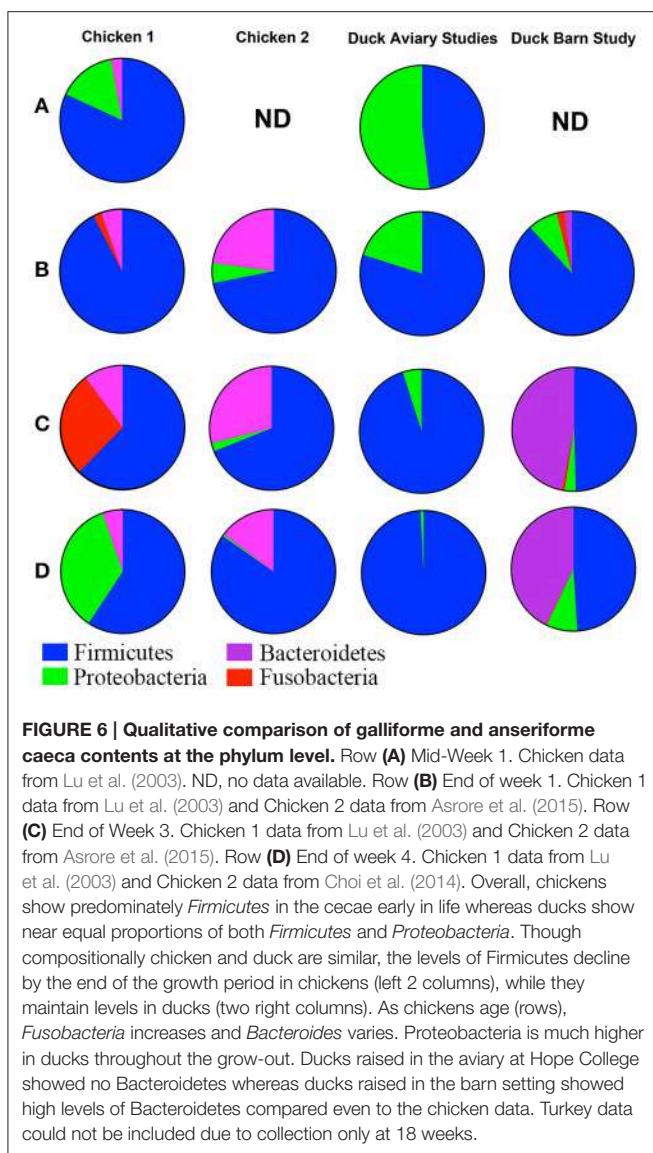
avian gastric motility, see Duke, 1982; Clench and Mathias, 1995). The internal villi allow for nutrient uptake while the blind end allows for increased content retention time and continuous reverse peristalsis to retain contents within the digestive tract (Shibata and Sogou, 1982). With decreased oxygen levels, this area becomes ideal for increased bacterial loads, aiding in digestion and uptake of crucial nutrients. The development of the digestive system of the Pekin duck occurs rapidly posthatch. In particular, it has been shown that the ileal and jejunal mucosa undergo large morphological changes during the first 7 days posthatch, including increases in villus height and crypt depth (Applegate et al., 1999, 2005). These changes coincide with drastic changes in the microbiome of the caecum from dominance by Proteobacteria to Firmicutes observed in this study. In comparison to galliforms, other areas of the digestive tract have been found to have less bacterial diversity possibly due to the constant motion of contents through the tract (Gong et al., 2007; Stanley et al., 2014). Even so, the filling of contents in the caecum from the small intestine suggest that these developmental changes could have significant impact on the microbiome and microenvironments found within the caecum.

Surprisingly, the phylum Bacteroidetes was not observed in any of the ducks from aviary studies 1 and 2 at significant levels. This group, to our knowledge, is always associated with gut microbiomes in other animals and has been linked to health states of individuals (Ley et al., 2008; Cho and Blaser, 2012; Tims et al., 2013). For example, shifts in Firmicutes/Bacteroidetes ratios have been observed in lean and obese model systems (Gilbert et al., 2016), as well as in broiler chickens and Pekin ducks based on fecal swabs and detection via qPCR (Angelakis and Raoult, 2010). Further, Bacteroidetes was found to be a dominant phylum present in the caecal contents of 12–14 week old Pekin ducks (Vasai et al., 2014). While both aviary studies were conducted in the same aviary setting, they were separated by 6 months in time. Bacteroidetes was identified in water samples from the aviary setting (data not shown), suggesting that the ducks were ingesting a potential source of this group of bacteria. However, this group of organisms did not establish a detectable population in the duck caecal cavity. Given that these are the first comprehensive data sets for the developing Pekin duck caecal microbiome, it was possible that the composition is very different from all other available data from animal gut environments. In a parallel study by our group (Schenk et al., 2016), we investigated the caecal contents of ducks raised in a commercial production environment that used well water and an open barn design, allowing us to compare the aviary and barn data sets to determine if the pattern of absence of Bacteroidetes as a major component of the microbiome of Pekin ducks was consistent across environments. These results suggest that the absence of Bacteroidetes in ducks raised in the aviary setting is dependent on local environmental factors. The aviary is kept clean throughout the time that ducks are present and sterilized bedding is used, whereas in a barn setting with thousands of ducks and open air flow, the ducks will be exposed to many more sources of environmental microbes. In all cases, caecal contents came from healthy, well developing individuals, which raises questions about how different microbial

populations in the caecum affect agriculturally important phenotypes.

The most common pathogen in anseriformes is *Riemerella anatipestifer* (RA) (Wobeser, 1997), which occurs globally in wild and commercially raised ducks (Brogden, 1989). RA went unexplored for quite some time, and has only recently been taxonomically classified and further characterized (Ryll et al., 2008). Its pathogenesis remains unknown, though we hypothesized that it might be part of the endogenous gut microbiome and that it may become opportunistic under the right environmental conditions. However, RA was not observed in any caecal contents in this study. The respiratory tract and the epidermis represent two other areas of the Pekin duck that could harbor a subpopulation of RA and are of interest in further studies. Sequences of rRNA genes of non-serotypable RA-like strains isolated from the pharyngeal flora of healthy Pekin ducks were found to be 99% identical to those of RA (Ryll et al., 2008). Other genera that include pathogens, such as *Streptococcus* and *Enterococcus* appear to be normal constituents of the duck caecal microbiome, comprising up to 50% of the population during development in some individuals and in later aged ducks (Vasai et al., 2014). The ducks in this study harboring *Streptococcus* were healthy, and this highlights the distinction between presence of a potential pathogen and actual instance of disease (Casadevall and Pirofski, 2014).

While there are many caveats associated with comparison of microbial population data across different studies (Stanley et al., 2013; Gilbert et al., 2016), we present a high level comparison of galliforme and anseriforme caecal contents to identify major differences that may be inherent to the two bird types. Data taken from publications that describe the microbiome of chickens (Lu et al., 2003; Choi et al., 2014; Asrore et al., 2015) and turkeys (Scupham et al., 2008) were used for the qualitative assessment of caecal contents among different commercial birds and are illustrated in **Figure 6**. The primary difference at a phylum level comparison is the persistence of Proteobacteria throughout the maturation period of the duck, whereas this phylum is a very low percentage of the population or undetectable in broilers after the earliest sampled time point of 3 days posthatch. Broiler caeca microbial populations are dominated by Firmicutes from an early age and persist through >40 days posthatch, comprising well over half of the population (Lu et al., 2003; Choi et al., 2014; Asrore et al., 2015). Turkey caeca microbial populations at 18 weeks of age, are comprised of 52% Bacteroidetes, 33% Firmicutes, 5% Proteobacteria, 4% Deferribacteres, and 6% unclassified bacteria (Scupham et al., 2008). Each of these taxa are represented in the barn-raised ducks at the most mature time point tested at 40% (Bacteroidetes), 46% (Firmicutes), 8% (Proteobacteria), and 0.6% (Defferibacteres). Additional phyla represented in the most mature age group for ducks include Tenericutes (0.7%) and Actinobacteria (4%) (**Figure 5**), neither of which were reported as present at more than a fraction of a percent in broilers and turkey. The differences within the galliforme genus and between galliformes and anseriformes may be the result of host genetics, environmental conditions, feed type, or immunity against species-specific pathogens. Various other factors may be at play causing differences between data



sets including farming practices, age, breed, and experimental design of studies. However, it has been shown that even within a single study of chickens that the microbiomes of the different study groups can vary significantly (Stanley et al., 2013). Turkeys included in this comparison were sacrificed at 18 weeks of age, which is much older than a 1- to 4-week old chick or duckling. This age component, in itself, may have elicited considerable observed differences in caecal contents among the avian species. However, it is apparent that the caecal contents of anseriformes and galliformes are different, consistent with known differences in physiology and development (Applegate et al., 2005). Thus, it is possible that differences in environment or in feed composition could have considerably different effects in galliforms compared to anseriforms.

Prebiotic and probiotic supplementation has been popular in commercial farming to increase resistance to disease, increase growth rates, and improve overall poultry health (Lee et al., 1999;

Roto et al., 2015). Ideal cocktails of probiotics are constantly being evaluated, singling out specific species that will prove to have the most positive effects. Understanding which bacteria are beneficial and how they promote health may improve methods to increase growth yields and decrease disease rates. Lactic acid bacteria such as *Lactobacillus acidophilus* have been shown to increase growth rates while decreasing *E. coli* production and are commonly used in commercial poultry feed (Watkins et al., 1982). *Lactobacillus salivarius* and *Lactobacillus agilis* have also been shown to increase growth yields in broiler chickens (Lan et al., 2003). Though widely used, probiotic effects and consequences *in vivo* are not extensively understood, especially in anseriformes. Members of Clostridia cluster IV, a group known to be associated with butyrate production and to be important for gut homeostasis in mammals (Lopetuso et al., 2013), were identified as significant members of the microbiome of ducks in this study. It has been shown that changes in the distribution of these group members, in particular *Oscillospira*, at the order level and below are associated with differences in body mass index (BMI) between pairs of monozygotic twins (Tims et al., 2013). *Butyrivibrio* is associated with butyrate production and is a development target for probiotics aimed at mitigating symptoms of irritable bowel syndrome (Geirnaert et al., 2015). These observations raise interesting possibilities for studying probiotic supplementation of duck feeding practices and for study of differential weight gain in ducks. Introduction of probiotics may induce changes in endogenous microbiome populations, possibly creating new outlets for disease expression and immune system alterations. Creating a reference point in healthy ducks not receiving dietary supplements for comparison to microbial caecal contents from treated ducks will allow for changes to be tracked and analyzed as health changes occur. However, the observation that major phylum level changes can occur with changes in the setting for a study (here, the absence of Bacteroidetes in Aviary Studies 1 and 2 compared to the Barn Study 3) dictates careful design of microbiome based studies to include internal controls rather than reliance on comparison between different studies. This is consistent with recently recommended best practices for microbiome experimental design (Goodrich et al., 2014; Westcott and Schloss, 2015).

In summary, microbial population succession correlated strongly with duck age, exhibiting a clear transition in dominant taxa as ducks matured. Caecal contents of ducklings showed high levels of Proteobacteria that decreased with age, but was maintained at a higher proportion of the population than seen in chicken or turkey. The taxonomic transition led to a dominance of Firmicutes for the remainder of the ducks' life span in an aviary setting. In contrast, the transition led to two major phyla, Firmicutes, and Bacteroidetes, in ducks raised in a barn setting. This taxonomic milieu proved to be much different than both broiler chicken and turkey gut microbiomes described previously (Lu et al., 2003; Scupham et al., 2008; Choi et al., 2014), whereas later time points in development assessed in this study are consistent with caecal microbiomes of 12–14 week old Pekin ducks (Vasai et al., 2014). *R. anatipestifer* was not found in the samples collected in either aviary or farm settings; other genera that contain common pathogens of anseriformes were

identified. Characterization of microbiomes reflective of overall healthy ducks will be used for further assessments of commercial production practices. In particular, these data will allow for investigation of the origin and development of pathogens in commercial flocks, evaluation, and development of prebiotics and probiotics, other practices that potentially improve growth yields, and maintenance of food safety.

AUTHOR CONTRIBUTIONS

AB, SF, and GS designed the study. AP, SF, and GF acquired and processed samples in aviary and barn environments. AB and AP performed data analyses and drafted the manuscript. All authors participated in data interpretation and manuscript editing.

FUNDING

The authors thank Maple Leaf Farms, Inc. for their support for this research project. This work was funded in part by

National Science Foundation DBI Award 1229585 to AB and by the National Science Foundation REU award 0754293 to GF (Co-PI).

ACKNOWLEDGMENTS

The authors thank Hope College students Allyson Schenk and Alexis Meelker for assistance with sample collection, the Department of Biology at Hope College for their continued support of our research, and Sarah Owens and Jack Gilbert of the Institute for Genomics and Systems Biology Next Generation Sequencing (IGSB-NGS) Core Facility at Argonne National Laboratory for assistance with sequencing and initial discussion of data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2016.02125/full#supplementary-material>

REFERENCES

- AGMRC (2012). *Ducks and Geese, Ag Marketing Resource Center*. Available online at: <http://www.agmrc.org/commodities-products/livestock/poultry/ducks-and-geese/>
- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106
- Angelakis, E., and Raoult, D. (2010). The increase of *Lactobacillus* species in the gut flora of newborn broiler chicks and ducks is associated with weight gain. *PLoS ONE* 5:e10463. doi: 10.1371/journal.pone.0010463
- Applegate, T. J., Karcher, D. M., and Lilburn, M. S. (2005). Comparative development of the small intestine in the turkey poult and Pekin duckling. *Poult. Sci.* 84, 426–431. doi: 10.1093/ps/84.3.426
- Applegate, T. J., Ladwig, E., Weisert, L., and Lilburn, M. S. (1999). Effect of hen age on intestinal development and glucose tolerance of the Pekin duckling. *Poult. Sci.* 78, 1485–1492.
- Asrore, S. M. M., Sieo, C. C., Chong, C. W., Gan, H. M., and Ho, Y. W. (2015). Deciphering chicken gut microbial dynamics based on high-throughput 16S rRNA metagenomics analyses. *Gut Pathog.* 7, 1–12. doi: 10.1186/s13099-015-0051-7
- Biddle, A., Stewart, L., Blanchard, J., and Leschine, S. (2013). Untangling the genetic basis of fibrolytic specialization by lachnospiraceae and ruminococcaceae in diverse gut communities. *Diversity* 5, 627–640. doi: 10.3390/d5030627
- Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi: 10.1038/nmeth.2276
- Brogden, K. A. (1989). “*Pasteurella* anatipesf infection,” in *Pasteurella and Pasteurellosis*, eds. C. Adlam and J. Rutter (London: Academic Press, Inc.), 115–129.
- Caporaso, J. G., Bittinger, K., Bushman, F. D., Desantis, T. Z., Andersen, G. L., and Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26, 266–267. doi: 10.1093/bioinformatics/btp636
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1–4. doi: 10.1038/ismej.2012.8
- Casadevall, A., and Pirofski, L. A. (2014). Ditch the term pathogen. *Nature* 516, 165–166. doi: 10.1038/516165a
- Cho, I., and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* 13, 260–270. doi: 10.1038/nrg3182
- Choi, J. H., Kim, G. B., and Cha, C. J. (2014). Spatial heterogeneity and stability of bacterial community in the gastrointestinal tracts of broiler chickens. *Poult. Sci.* 93, 1942–1950. doi: 10.3382/ps.2014-03974
- Clench, M. H., and Mathias, J. R. (1995). The avian cecum: a review. *Wilson Bull.* 107, 93–121. Available online at: <http://www.jstor.org/stable/4163516>
- D’Argenio, V., and Salvatore, F. (2015). The role of the gut microbiome in the healthy adult status. *Clin. Chim. Acta* 451, 97–102. doi: 10.1016/j.cca.2015.01.003
- DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimeric-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05
- Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* 14, 927–930. doi: 10.1658/1100-9233(2003)014[0927:VAPORF]2.0.CO;2
- Duke, G. E. (1982). Gastrointestinal motility and its regulation. *Poult. Sci.* 61, 1245–1256.
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461
- Eren, A. M., Sogin, M. L., Morrison, H. G., Vineis, J. H., Fisher, J. C., Newton, R. J., et al. (2014). A single genus in the gut microbiome reflects host preference and specificity. *ISME J.* 9, 1–11. doi: 10.1038/ismej.2014.97
- Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P., and Forano, E. (2012). Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* 3, 289–306. doi: 10.4161/gmic.19897
- Frank, D. N., St Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13780–13785. doi: 10.1073/pnas.0706625104
- Funkhouser, L. J., and Bordenstein, S. R. (2013). Mom knows best: the universality of maternal microbial transmission. *PLoS Biol.* 11:e1001631. doi: 10.1371/journal.pbio.1001631
- Gabriel, I., Lessire, M., Mallet, S., and Guillot, J. F. (2006). Microflora of the digestive tract: critical factors and consequences for poultry. *Worlds Poult. Sci. J.* 62, 499–511. doi: 10.1079/WPS2006111
- Geirnaert, A., Wang, J., Tinck, M., Steyaert, A., Van den Abbeele, P., Eeckhaut, V., et al. (2015). Interindividual differences in response to treatment with

- butyrate-producing *Butyrivibrio pullicaecorum* 25-3T studied in an *in vitro* gut model. *FEMS Microbiol. Ecol.* 91:fiv054. doi: 10.1093/femsec/fiv054.
- Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., et al. (2014). The treatment-naïve microbiome in new-onset Crohn's disease. *Cell Host Microbe* 15, 382–392. doi: 10.1016/j.chom.2014.02.005
- Gilbert, J. A., Quinn, R. A., Debelius, J., Xu, Z. Z., Morton, J., Garg, N., et al. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* 535, 94–103. doi: 10.1038/nature18850
- Gong, J., Si, W., Forster, R. J., Huang, R., Yu, H., Yin, Y., et al. (2007). 16S rRNA gene-based analysis of mucosa-associated bacterial community and phylogeny in the chicken gastrointestinal tracts: from crops to ceca. *FEMS Microbiol. Ecol.* 59, 147–157. doi: 10.1111/j.1574-6941.2006.00193.x
- Goodrich, J. K., Di Rienzi, S. C., Poole, A. C., Koren, O., Walters, W. A., Caporaso, J. G., et al. (2014). Conducting a microbiome study. *Cell* 158, 250–262. doi: 10.1016/j.cell.2014.06.037
- Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504. doi: 10.1101/gr.112730.110
- Lan, P. T. N., Binh, L. T., and Benno, Y. (2003). Impact of two probiotic *Lactobacillus* strains feeding on fecal lactobacilli and weight gains in chicken. *J. Gen. Appl. Microbiol.* 49, 29–36. doi: 10.2323/jgam.49.29
- Lee, Y. K., Nomoto, K., Salminen, S., and Gorbach, S. (1999). *Handbook of Probiotics*. New York, NY: Wiley Interscience.
- Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., et al. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric crohn's disease. *Cell Host Microbe* 18, 489–500. doi: 10.1016/j.chom.2015.09.008
- Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., et al. (2008). Evolution of mammals and their gut microbes. *Science* 320, 1647–1651. doi: 10.1126/science.1155725
- Lopetuso, L. R., Scaldaferri, F., Petito, V., and Gasbarrini, A. (2013). Commensal clostridia: leading players in the maintenance of gut homeostasis. *Gut Pathog.* 5:23. doi: 10.1186/1757-4749-5-23
- Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005
- Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. *ISME J.* 5, 169–172. doi: 10.1038/ismej.2010.133
- Lu, J., Idris, U., Harmon, B., Hofacre, C., Maurer, J. J., and Lee, M. D. (2003). Diversity and succession of the intestinal bacterial community of the maturing broiler chicken. *Appl. Environ. Microbiol.* 69, 6816–6824. doi: 10.1128/AEM.69.11.6816-6824.2003
- McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139
- McGlone, J., Swanson, J., Ford, S., Mitloehner, F., Grandin, T., Ruegg, P., et al. (2010). *Guide for the Care and Use of Agricultural Animals in Research and Teaching*, 3rd Edn. Champaign, IL: Federation of Animal Science Societies. Available online at: <http://www.fass.org>
- McMurdie, P. J., and Holmes, S. (2013). Phyloseq: an R Package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8:e61217. doi: 10.1371/journal.pone.0061217
- Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., et al. (2013). "Advancing our understanding of the human microbiome using QIIME," in *Methods in Enzymology*, ed. E. F. Delong (New York, NY: Elsevier Inc.), 371–444. doi: 10.1016/B978-0-12-407863-5.00019-8
- Nelson, M. C., Morrison, H. G., Benjamin, J., Grim, S. L., and Graf, J. (2014). Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS ONE* 9:e94249. doi: 10.1371/journal.pone.0094249
- Neo, E., La, T., Phillips, N. D., Alikani, M. Y., and Hampson, D. J. (2013). The pathogenic intestinal spirochaete *Brachyspira pilosicoli* forms a diverse recombinant species demonstrating some local clustering of related strains and potential for zoonotic spread. *Gut Pathog.* 5:24. doi: 10.1186/1757-4749-5-24
- Nicholson, J. K., Holmes, E., and Wilson, I. D. (2005). Gut microorganisms, mammalian metabolism and personalized health care. *Nat. Rev. Micro.* 3, 431–438. doi: 10.1038/nrmicro1152
- Van Opstal, E. J. V., and Bordenstein, S. R. (2015). Rethinking heritability of the microbiome. *Science* 349, 1172–1173. doi: 10.1126/science.aab3958
- Pérez de Rozas, A. M. (2004). "A comparative study of intestinal microbial diversity from birds, pigs and rabbits by Restriction Fragment Length Polymorphism analysis," in *Reproduction Nutrition Development*, eds J.-C. Thiéry, P. Guesnet, and M. Guillotot (London: EDP Sciences), S4.
- Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490
- Reid, G., Younes, J. A., van der Mei, H. C., Gloor, G. B., Knight, R., and Busscher, H. J. (2011). Microbiota restoration: natural and supplemented recovery of human microbial communities. *Nat. Rev. Microbiol.* 9, 27–38. doi: 10.1038/nrmicro2473
- Roto, S. M., Rubinelli, P. M., and Ricke, S. C. (2015). An introduction to the avian gut microbiota and the effects of yeast-based prebiotic-type compounds as potential feed additives. *Front. Vet. Sci.* 2:28. doi: 10.3389/fvets.2015.00028
- Ryll, M., Christensen, H., Bisgaard, M., Christensen, J. P., Hinz, K. H., and Köhler, B. (2008). Studies on the prevalence of *riemerella anatum* in the upper respiratory tract of clinically healthy ducklings and characterization of untypable strains. *J. Vet. Med. Ser. B* 48, 537–546. doi: 10.1111/j.1439-0450.2001.00471.x
- Schenk, A., Porter, A. L., Alenciks, E., Frazier, K., Best, A. A., Fraley, S. M., et al. (2016). Increased water contamination and grow-out Pekin duck mortality when raised with water troughs compared to pin-metered water lines using a United States management system. *Poult. Sci.* 95, 736–748. doi: 10.3382/ps/pev281
- Scupham, A. J., Patton, T. G., Bent, E., Bayles, D. O., Scupham, A. J., Patton, T. G., et al. (2008). Comparison of the cecal microbiota of domestic and wild turkeys. *Microb. Ecol.* 56, 322–331. doi: 10.1007/s00248-007-9349-4
- Shibata, H., and Sogou, M. (1982). [Gastrointestinal transit in the chicken using 198Au-colloid as a marker (author's transl)]. *Radioisotopes* 31, 82–87.
- Soergel, D. A. W., Dey, N., Knight, R., and Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J.* 6, 1440–1444. doi: 10.1038/ismej.2011.208
- Stanley, D., Geier, M. S., Hughes, R. J., Denman, S. E., and Moore, R. J. (2013). Highly variable microbiota development in the chicken gastrointestinal tract. *PLoS ONE* 8, e64290. doi: 10.1371/journal.pone.0084290
- Stanley, D., Hughes, R. J., and Moore, R. J. (2014). Microbiota of the chicken gastrointestinal tract: influence on health, productivity and disease. *Appl. Microbiol. Biotechnol.* 98, 4301–4310. doi: 10.1007/s00253-014-5646-2
- Tims, S., Derom, C., Jonkers, D. M., Vlietinck, R., Saris, W. H., Kleerebezem, M., et al. (2013). Microbiota conservation and BMI signatures in adult monozygotic twins. *ISME J.* 7, 707–717. doi: 10.1038/ismej.2012.146
- Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031. doi: 10.1038/nature05414
- van der Wielen, P. W., Keuzenkamp, D. A., Lipman, L. J., Van Knapen, F., and Biesterveld, S. (2002). Spatial and temporal variation of the intestinal bacterial community in commercially raised broiler chickens during growth. *Microb. Ecol.* 44, 286–293. doi: 10.1007/s00248-002-2015-y
- Van Immerseel, F., De Buck, J., Pasmans, F., Huyghebaert, G., Haesebrouck, F., Ducatelle, R., et al. (2004). Clostridium perfringens in poultry: an emerging threat for animal and public health. *Avian Pathol.* 33, 537–549. doi: 10.1080/03079450400013162
- Vasai, F., Brugirard Ricaud, K., Bernadet, M. D., Cauquil, L., Bouchez, O., Combes, S., et al. (2014). Overfeeding and genetics affect the composition of intestinal microbiota in *Anas platyrhynchos* (Pekin) and *Cairina moschata* (Muscovy) ducks. *FEMS Microbiol. Ecol.* 87, 204–216. doi: 10.1111/1574-6941.12217
- Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A., and Knight, R. (2013). EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* 2:16. doi: 10.1186/2047-217X-2-16
- Waite, D. W., and Taylor, M. W. (2014). Characterizing the avian gut microbiota: membership, driving influences, and potential function. *Front. Microbiol.* 5:223. doi: 10.3389/fmicb.2014.00223

- Watkins, B. A., Miller, B. F., and Neil, D. H. (1982). *In vivo* inhibitory effects of *Lactobacillus acidophilus* against pathogenic *Escherichia coli* in gnotobiotic chicks. *Poult. Sci.* 61, 1298–1308. doi: 10.3382/ps.0611298
- Wei, S., Morrison, M., and Yu, Z. (2013). Bacterial census of poultry intestinal microbiome. *Poult. Sci.* 92, 671–683. doi: 10.3382/ps.2012-02822
- Westcott, S. L., and Schloss, P. D. (2015). *De novo* clustering methods out-perform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 3:e1487. doi: 10.7717/peerj.1487
- Wobeser, G. A. (1997). *Diseases of Wild Waterfowl*. Boston, MA: Springer US.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Best, Porter, Fraley and Fraley. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The Importance of Bacterial Culture to Food Microbiology in the Age of Genomics

Alexander Gill*

Health Canada, Bureau of Microbial Hazards, Ottawa, ON, Canada

Culture-based and genomics methods provide different insights into the nature and behavior of bacteria. Maximizing the usefulness of both approaches requires recognizing their limitations and employing them appropriately. Genomic analysis excels at identifying bacteria and establishing the relatedness of isolates. Culture-based methods remain necessary for detection and enumeration, to determine viability, and to validate phenotype predictions made on the basis of genomic analysis. The purpose of this short paper is to discuss the application of culture-based analysis and genomics to the questions food microbiologists routinely need to ask regarding bacteria to ensure the safety of food and its economic production and distribution. To address these issues appropriate tools are required for the detection and enumeration of specific bacterial populations and the characterization of isolates for, identification, phylogenetics, and phenotype prediction.

Keywords: bacteria, culture, genomics, food, detection, characterization, subtyping

OPEN ACCESS

Edited by:

Sandra Torriani,
University of Verona, Italy

Reviewed by:

Eelco Franz,
Centre for Infectious Disease Control,
Netherlands

David Rodriguez-Lazaro,
University of Burgos, Spain
Johannes F. Imhoff,
GEOMAR Helmholtz Centre for Ocean
Research Kiel (HZ), Germany

***Correspondence:**

Alexander Gill
alex.gill@hc-sc.gc.ca

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 23 January 2017

Accepted: 18 April 2017

Published: 01 May 2017

Citation:

Gill A (2017) The Importance of Bacterial Culture to Food Microbiology in the Age of Genomics.
Front. Microbiol. 8:777.
doi: 10.3389/fmicb.2017.00777

INTRODUCTION

Genomics, the study of the encoding, structure and function of genetic information, can be considered to have emerged as a recognized discipline with the initial publication of the eponymous journal in 1987 (McKusick and Ruddle, 1987). Initial efforts to sequence the complete genome of organisms required heroic investments of ingenuity and resources. The first prokaryote sequence, that of *Haemophilus influenzae*, was published in 1995 (Fleischmann et al., 1995) and the first eukaryote sequence, *Saccharomyces cerevisiae*, in 1996 (Goffeau et al., 1996). The first human genome was completed in 2004, following a multinational effort with an estimated cost of 3 billion dollars (Schmutz et al., 2004). The return on these investments has been techniques and tools which have dramatically lowered the time and cost of sequencing and the time and expertise required for analysis. For bacteriologists, genomic analysis is becoming a routine tool, with whole genome sequencing (WGS) available affordably, in a matter of days and with analysis supported by online platforms (Jackson et al., 2016; Kwong et al., 2016; Lindsey et al., 2016; Whiteside et al., 2016; Yoshida et al., 2016).

The foundation of bacteriology as an experimental science was the development in the 19th century of cultural techniques using solid or liquid media. Culture allowed viable bacteria to be detected and isolated, facilitated the observation of metabolic activity, and provided biomass for further analysis. As genomic analysis of bacteria becomes available as a routine tool, there is a risk of a perception developing that genomic analysis of bacteria is superior or may supersede bacterial cultural methods due to the speed of analysis and quantity of data produced. The purpose of this paper is to discuss the questions that food microbiologists routinely consider regarding bacteria,

and to examine the application of cultural and genomic approaches to answering them. In food microbiology bacteria as infectious pathogens and toxin producers are a safety concern. Bacteria are also an economic concern, as their metabolic activity may enhance the economic value of foods or negatively impact it by altering sensory qualities or nutritional content.

Consideration of the strengths and limitations of cultural and genomic based methods of analysis of bacteria indicates that they provide different insights into the nature and behavior of bacteria, with the former revealing phenotypic characteristics and behavior and the latter genotypic information. Neither type of information is superior to the other, but rather they should be viewed as specialized and complementary tools which are suited to answering different experimental questions.

THE DETECTION AND ENUMERATION OF BACTERIA IN FOODS

The most commonly used forms of bacteriological analysis in food microbiology are detection and enumeration. The presence of specific bacteria and their concentration must be determined, to assess and control safety hazards, the potential for spoilage or to ensure correct product characteristics. The bacteria of interest to food microbiology can be divided into infectious agents, causes of foodborne intoxication, spoilage, and processing aids (**Table 1**). Metabolic activity of a bacterium may be considered as causing spoilage or as a processing aid depending upon the desirability of the changes that result.

Detection of specific types of bacteria can be achieved by cultural isolation, or by indicators such as biomolecules specific to the organism (e.g., nucleic acid sequences, antigens, toxins) or products of metabolism (e.g., gas, acid, substrates with chromogenic products) (Gill et al., 2014). For enumeration cell-concentration can be estimated by partitioning the sample upon a solid surface (e.g., agar media, membrane), between liquid aliquots (e.g., most probable number) or through direct

or indirect measures of biomass (e.g., optical density, Limulus amebocyte lysate assay).

Culture independent diagnostic platforms for infectious agents have been successfully commercialized in the health care sector (Caliendo et al., 2013; Zumla et al., 2014; Langley et al., 2015) and the potential for speed and automation has stimulated interest in the application of similar systems for food analysis (Anonymous, 2016a,b; Wang and Salazar, 2016). Platforms developed for health care cannot be easily adopted for use in food microbiology as analysis is significantly more challenging. Food microbiology samples are considerably more variable in type, heterogeneous in composition and the concentrations of target bacteria can be much lower. Also with the exception of stools, body fluids and tissue samples can normally be expected to contain a negligible microbiota.

The presence of non-target microbiota is a particular problem when testing for pathogens as closely related non-pathogens may result in false positives, which can have serious implications for producers. For example, the US Department of Agriculture (USDA, 2016) requires testing of raw ground beef components for shiga toxin-producing *E. coli* (STEC) which possess three traits: the virulence genes *stx* and *eae*, and six O-types considered of high risk. *E. coli* strains which possess one or two of these traits are considered of low risk and do not require the same regulatory response. The three traits, however, are not genetically linked and may be present within the sample in multiple different organisms (Delannoy et al., 2016).

Genomic technologies are considered appealing for culture-independent detection as reliability can be provided by the parallel detection of multiple genes or their transcription products. Though not currently possible, in principle, WGS using sufficiently long reads could detect and confirm the presence of the complete genome of multiple target organisms in a complex mix of DNA. In spite of accelerating advances, however, genomic technologies are not suitable for addressing two fundamental challenges in detection and enumeration; sensitivity and the determination of viability.

TABLE 1 | Examples of bacteria of concern to food microbiology.

Foodborne infectious agents	Foodborne intoxicants	Spoilage	Processing
<i>Brucella</i>	<i>Bacillus cereus</i>	<i>Acinetobacter</i>	Lactic Acid
<i>Campylobacter</i>	<i>Clostridum botulinum</i>	<i>Alcaligenes</i>	Bacteria
<i>Clostridium botulinum</i>	<i>Clostridium perfringens</i>	<i>Bacillus</i>	(<i>Lactobacillus</i> ,
<i>Clostridium perfringens</i>	<i>Staphylococcus aureus</i>	<i>Brochothrix</i>	<i>Lactococcus</i> ,
<i>Cronobacter</i>		<i>thermosphacta</i>	<i>Pediococcus</i> ,
pathogenic <i>Escherichia coli</i>		<i>Clostridium</i>	<i>Leuconostoc</i> ,
<i>Shigella</i>		<i>Cornebacterium</i>	<i>Streptococcus</i>)
<i>Salmonella enterica</i>		<i>Enterobacteriaceae</i>	
<i>Yersinia enterocolitica</i>		<i>Erwinia carotovora</i>	
<i>Listeria monocytogenes</i>		Lactic Acid Bacteria	
<i>Mycobacterium</i>		<i>Moraxellaceae</i>	
<i>Vibrio</i>		<i>Pseudomonas</i>	
		<i>Shewanella</i>	
		<i>putrefaciens</i>	
		<i>Vibrio</i>	

The Sensitivity of Cultural Methods for Bacterial Detection

The sensitivity or limit of detection (LOD) of methods of analysis for bacterial cells is the minimum concentration of cells that can be detected. Analysis for the presence of bacteria that cause foodborne intoxication, spoilage or serve as production aids does not generally require limits of detection below 100 CFU/g or ml. Spoilage and processing bacteria do not impact the quality of a product until they exceed a significant concentration, for example spoilage of red meats by *Pseudomonads* becomes apparent above 6 log CFU/cm² (Gill and Newton, 1980). Bacteria which causes intoxication need to reach relatively high concentrations in foods before significant toxin production occurs. For *C. botulinum* the threshold is 3 log CFU/g (Austin et al., 1998) and for *B. cereus* and *S. aureus* cell concentration must exceed 5 log CFU/g (FDA, 2012). However, some infectious agents have infectious doses estimated in the range of 10–100 cells (Todd et al., 2008), and the concentration of pathogen cells in outbreak associated products may be below 1 cell per 25 g (Gill and Oudit, 2015; Gill and Huszczynski, 2016). Thus, regulatory compliance testing of foods for infectious bacterial pathogens requires LODs approaching 1 cell per analytical unit, with analytical units of 10 g to 325 g depending on the specific pathogen and food (FDA, 2016; Health Canada, 2016; USDA, 2016).

Without enrichment, no existing technologies can approach this sensitivity (Wang and Salazar, 2016). Whether or not analysis is based on detection of cells or biomolecules, the target of analysis needs to be separated from the surrounding complex organic matrix, without loss of the target by adherence to the analytical apparatus. The relatively large analytical unit sizes (10 g to 325 g) make it impractical to assay anything other than a smaller aliquot of the analytic unit (0.1 to 1 ml) and many foods are composed of solids, gels and suspensions, with consequent heterogeneous distribution of bacterial cells. A method of analysis which is dependent upon the probability of the target being present in an aliquot of the analytical sample is inherently unreliable.

Cultural enrichment resolves the challenge of high sensitivity bacterial detection by amplifying the analytic target to raise the concentration and distribute it homogeneously through an aqueous suspension. This ensures that detection is no longer a probabilistic process and sample handling is greatly eased. The only limitation is that the minimum time required for sample analysis is determined by the enrichment period. This will be determined by the time required for cells to begin replication (repair injury and exit lag phase) and the time required to reach the LOD of the method of analysis (growth rate). The enrichment period could be reduced by a concentration process once the analytic target is homogeneously distributed in the suspension. However, the time and resources required for concentration may make this less efficient than extending the enrichment period.

Determination of Bacterial Viability

For the bacteriological analysis of foods, it is highly desirable that the method of analysis does not confound viable cells (cells with the potential to replicate), with non-viable cells or cell debris.

Foods often undergo processing steps which impact bacterial survival: the resulting population of cells may include viable cells, cells that can replicate following repair (reversibly injured) and cells which can not replicate but retain metabolic activity (irreversibly injured) (Wu, 2008). Cells of infectious agents that cannot replicate pose no threat to health. Non-replicating toxin producers only pose a threat if their concentration is already high enough to present a risk. Similarly, the presence of non-replicating cells of spoilage and processing bacteria are of no relevance to food quality.

Analytical methods which detect the presence of biomolecules, such as DNA, RNA, or proteins, cannot determine whether those biomolecules represent viable cells or not. Cell replication is a complex process, in which multiple regulatory mechanisms must coordinate the synthesis and localization of a vast array of structural and functional molecules (Reyes-Lamothe et al., 2012; Murray and Koh, 2014; Murray, 2016). The failure of the cell to complete any essential function can stall cell growth and division. Confirmation of the presence of any single essential cell component does not exclude other deficiencies that would inhibit cell replication. Thus, viability can only be determined by two methods, determination of the presence and functionality of all the molecules required, or simply waiting for the cell to exit lag phase and allowing replication to occur.

Assessment of viability may be further complicated by the potential for vegetative bacterial cells, including some foodborne pathogens, to enter into an alternate physiological state, viable-but-nonculturable (VBNC). The VBNC state can be triggered by a variety of physiochemical stresses, with cells ceasing replication but continuing metabolic activity (Pinto et al., 2015). Experimentally distinguishing VNBC states from injury is experimentally complicated, but since by definition VNBC cells can resume replication following exposure to an appropriate resuscitation stimulus (Pinto et al., 2015) identifying the correct stimulus for resuscitation rather than abandoning culture appears a more productive response.

CHARACTERING BACTERIA

Bacterial isolates are characterized to confirm identity, to establish relationships between isolates and to understand the behavior (phenotype) of the bacterium. Though a variety of cultural and molecular methods for characterization are available genomic analysis is superseding them. Genomics may have clear superiority over other approaches in identifying and subtyping isolates, but its application to predict phenotype is much less reliable.

Identification and Subtyping

Genomic analysis to identify isolates to the genus or species level by 16S rRNA sequence and the detection of specific gene markers by polymerase chain reaction based methods is more reliable and has greater discrimination than phenotypic methods, particularly for identification below the species level (Pace, 2009; Yilmaz et al., 2014). WGS analysis is superseding other approaches, as though complex computational analysis is required, a single wet lab

process can provide information on the presence of multiple gene markers and phylogenetic relationship. Additionally, sequence data can be retrospectively analyzed for additional markers or potential relationships (Franz et al., 2016; Ronholm et al., 2016).

Establishing phylogenetic relationships for bacterial isolates below the species level can link isolates from clinical, food and environmental samples, for outbreak identification and source tracking (Fu and Li, 2014). Single polynucleotide polymorphism (SNP) analysis of WGS data can provide unprecedented discrimination between isolates (Holt et al., 2008; Chin et al., 2011). Methods for the prediction from WGS data of established genetic subtyping such as pulsed field gel electrophoresis, multilocus sequence typing and multiple-locus variable number tandem repeat analysis are being developed, but the accuracy can be compromised by short read lengths (Kwong et al., 2016; Yoshida et al., 2016). Whether subtyping of isolates is by SNP or alternatives, the only limitation on the relating isolates is the comprehensiveness of WGS and accompanying metadata available. For example, WGS data has been used to deduce the geographical origin of isolates (Weedmark et al., 2015; Hoffmann et al., 2016).

It should be noted that the application of WGS data for isolate identification, subtyping and source tracking in the context of public health, regulatory, and commercial decision making is very recent and standards for analysis and interpretation have yet to be established. Reported results are dependent upon the sequencing platform used (length of reads, error rate, genome coverage), and the data analysis pipeline (Pettengill et al., 2014; Wang et al., 2015). Data interpretation may be significantly affected by choices such as nucleotide identification algorithms, assembly method (*de novo* or reference guided) and whether the shared or core genome of isolates is compared. The need to address these issues by establishing analytical standards is recognized, but there will be a transition period until a consensus on standards emerges (Franz et al., 2016; Ronholm et al., 2016).

Predicting Bacterial Phenotype

The complementary nature of genomic and phenotypic data is apparent when attempting to understand and predict bacterial behavior. There is wide range of bacterial behavior of interest to the food microbiologist. These include the potential for survival and replication during food production and distribution, spoilage potential, and the hazard potential of pathogens. Genomic analysis allows researchers to rapidly detect known genes, putative genes, and other defined features of the genome. However, relating genotype to phenotype with accuracy is highly challenging. Many phenotypic characteristics are the product of multiple genes and their regulatory systems. Current knowledge of any but a handful of biological regulatory systems is far from perfect. The same phenotype may result from multiple mechanisms with differing genotypes (Wilson, 2014). The phenotype may also be dependent upon interactions with other organisms (Sanchez-Vizuete et al., 2015; Chanos and Mygind, 2016). Genetically homogenous cells may be phenotypically heterogeneous, as observed in persister cells and biofilms (Grote et al., 2015; Van Acker and Coenye, 2016; Verstraeten et al., 2016). Epigenetic inheritance, DNA methylation (Adhikari and

Curtis, 2016) and small RNAs (Houry-Zeevi and Rechavi, 2016) have been identified as playing a role in determining bacterial phenotype, but the mechanisms are not well understood. The significance of other potential epigenetic mechanisms such as prions (Pallarès et al., 2015), self-sustaining metabolic loops, and structural templating of membranes in bacteria is unknown (Jablonka and Lamb, 2005).

The challenges and the opportunities presented by predicting phenotype from genotype are illustrated by the prediction of antimicrobial resistance (AMR) from WGS data. Many WGS analysis platforms provide output of AMR associated genes, but the utility of this information is questionable. A 2016 review of the potential for AMR prediction by WGS conducted for the European Committee on Antimicrobial Susceptibility Testing concluded that, for the purposes of informing clinical decisions, "The published evidence for using WGS as a tool to infer antimicrobial susceptibility accurately is currently either poor or non-existent" (Ellington et al., 2016). Though prediction accuracy may be limited, the ability to rapidly screen large populations of strains for a potential phenotype is still useful. Knowles et al. (2016) used WGS data to determine whether a specific STEC strain possessed AMR genes that were relatively uncommon among *E. coli* and determined the presence of trimethoprim resistance genes. Trimethoprim resistance was confirmed experimentally and the addition of trimethoprim to enrichment broth was demonstrated to aid isolation (Knowles et al., 2016). In an outbreak investigation, this approach could be used in the analysis of foods for strains previously isolated from patients.

The purpose of discussing the limitations of genomics is not to imply that the application of genomics to answer these questions is inappropriate. When complemented with phenotypic data and studies of physiological mechanisms, genomic data is a powerful tool to improve our understanding, but the challenges in relating genomic data to phenotype must be recognized. Decision making related to food safety or food processing should not be made solely on the basis of genomic data, but needs to be supported by phenotypic data, which in turn require culture. When grounded in phenotypic data, genomic data has the potential to enhance culture methods (Knowles et al., 2016) or to develop culture method for previously unculturable organisms (Renesto et al., 2003). As databases of WGS data expand, genomic analysis can be used to rapidly screen large populations for strains that potentially possess desired phenotypes, or to select experimental strains that are representative of a larger population.

CONCLUSION

Genomic technologies are tools with enormous potential for increasing our understanding of bacteria and solving practical problems in food microbiology, but like any tools the benefits and costs are dependent upon how we choose to employ them. Medical microbiology is faced with a set of unnecessary challenges due to the trend of abandoning cultural isolation for culture-independent diagnostic testing. At the clinical level this presents difficulties distinguishing viable from non-viable

organisms and in data interpretation when multiple organisms are present. At the public health level this results in the inability to collect epidemiological data such as, subtype and AMR, and prohibits further characterization and research (Janda and Abbott, 2014; Huang et al., 2016). Food microbiologists should learn from this example and consider how to maximize the benefits without losing the advantages of alternate technologies. Genomic analysis is becoming the standard method for the identification and phylogenetics of bacteria, but culture remains necessary to achieve the required sensitivity of detection and enumeration and to determine viability. Just as crucially, culture is needed to provide the isolates with which to conduct experiments to test hypotheses generated from genomic data. When phenotype is predicted from genomic data we are creating

a model of a biological system and the great value of such models as noted by Jeremy Gunawardena (2014) is to reveal the limitations of our understanding.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor to this work and approved it for publication.

FUNDING

The author's research was supported by Health Canada.

REFERENCES

- Adhikari, S., and Curtis, P. D. (2016). DNA methyltransferases and epigenetic regulation in bacteria. *FEMS Microbiol. Rev.* 40, 575–591. doi: 10.1093/femsre/fwu023
- Anonymous (2016a). Available at: <https://www.mitacs.ca/en/projects/development-rapid-point-care-test-food-water-safety> [accessed November 29, 2016].
- Anonymous (2016b). Available at: <http://www.newswire.ca/news-releases/funding-awarded-to-detect-e coli-in-food-processing-facilities-512717881.html> [Accessed November 29, 2016].
- Austin, J. W., Dodds, K. L., Blanchfield, B., and Farber, J. M. (1998). Growth and toxin production by *Clostridium botulinum* on inoculated fresh-cut packaged vegetables. *J. Food Prot.* 61, 324–328. doi: 10.4315/0362-028X-61-3.324
- Caliendo, A. M., Gilbert, D. N., Ginocchio, C. C., Hanson, K. E., May, L., Quinn, T. C., et al. (2013). Better tests, better care: improved diagnostics for infectious diseases. *Clin. Infect. Dis.* 57, S139–S170. doi: 10.1093/cid/cit578
- Chanos, P., and Mygind, T. (2016). Co-culture-inducible bacteriocin production in lactic acid bacteria. *Appl. Microbiol. Biotechnol.* 100, 4297–4308. doi: 10.1007/s00253-016-7486-8
- Chin, C. S., Sorenson, J., Harris, J. B., Robins, W. P., Charles, R. C., Jean-Charles, R. R., et al. (2011). The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* 364, 33–42. doi: 10.1056/NEJMoa1012928
- Delannoy, S., Chaves, B. D., Ison, S. A., Webb, H. E., Beutin, L., Delaval, J., et al. (2016). Revisiting the STEC testing approach: using espK and espV to make enterohemorrhagic *Escherichia coli* (EHEC) detection more reliable in beef. *Front. Microbiol.* 7:1. doi: 10.3389/fmicb.2016.00001
- Ellington, M. J., Ekelund, O., Aarestrup, F. M., Canton, R., Doumith, M., Giske, C., et al. (2016). The role of whole genome sequencing (WGS) in antimicrobial susceptibility testing of bacteria: report from the EUCAST subcommittee. *Clin. Microbiol. Infect.* 23, 2–22. doi: 10.1016/j.cmi.2016.11.012
- FDA (2012). *Bad Bug Book: Foodborne Pathogenic Microorganisms and Natural Toxins*, 2nd Edn. Available at: <http://www.fda.gov/downloads/Food/FoodborneIllnessContaminants/UCM297627.pdf> [accessed November 24, 2016].
- FDA (2016). *Bacteriological Analytical Manual*. Available at: <http://www.fda.gov/Food/FoodScienceResearch/LaboratoryMethods/ucm2006949.htm> [accessed November 24, 2016].
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512. doi: 10.1126/science.7542800
- Franz, E., Gras, L. M., and Dallman, T. (2016). Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. *Curr. Opin. Food Sci.* 8, 74–79. doi: 10.1016/j.cofs.2016.04.004
- Fu, L.-L., and Li, J.-R. (2014). Microbial source tracking: a tool for identifying sources of microbial contamination in the food chain. *Crit. Rev. Food Sci. Nutr.* 54, 699–707. doi: 10.1080/10408398.2011.605231
- Gill, A., and Huszcynski, G. (2016). Enumeration of *Escherichia coli* O157:H7 in outbreak-associated beef patties. *J. Food Prot.* 79, 1266–1268. doi: 10.4315/0362-028X.JFP-15-521
- Gill, A., and Oudit, D. (2015). Enumeration of *Escherichia coli* O157 in outbreak-associated Gouda cheese made with raw milk. *J. Food Prot.* 78, 1733–1737. doi: 10.4315/0362-028X.JFP-15-036
- Gill, A. O., Greer, G. G., and Nattress, F. M. (2014). "Microbiological analysis: standard methods," in *Encyclopedia of Meat Sciences*, Vol. 2, 2nd Edn, eds C. Devine and M. Dikeman (Oxford: Elsevier), 306–316.
- Gill, C. O., and Newton, K. G. (1980). Development of bacterial spoilage at adipose tissue surfaces of fresh meat. *Appl. Environ. Microbiol.* 39, 1076–1077.
- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science* 274, 563–567. doi: 10.1126/science.274.5287.546
- Grote, J., Krysciak, D., and Streit, W. R. (2015). Phenotypic heterogeneity, a phenomenon that may explain why quorum sensing does not always result in truly homogenous cell behavior. *Appl. Environ. Microbiol.* 81, 5280–5289. doi: 10.1128/AEM.00900-15
- Gunawardena, J. (2014). Models in biology 'accurate descriptions of our pathetic thinking'. *BMC Biol.* 12:29. doi: 10.1186/1741-7007-12-29
- Health Canada (2016). *The Compendium of Analytical Methods*. Available at: <http://www.hc-sc.gc.ca/fn-an/res-rech/analy-meth/microbio/index-eng.php> [accessed November 24, 2016].
- Hoffmann, M., Luo, Y., Monday, S. R., Gonzalez-Escalona, N., Ottesen, A. R., Muruvanda, T., et al. (2016). Tracing origins of the *Salmonella* Bareilly strain causing a food-borne outbreak in the United States. *J. Infect. Dis.* 213, 502–508. doi: 10.1093/infdis/jiv297
- Holt, K. E., Parkhill, J., Mazzoni, C. J., Roumagnac, P., Weill, F. X., Goodhead, I., et al. (2008). High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat. Genet.* 40, 987–993. doi: 10.1038/ng.195
- Houri-Zeevi, L., and Rechavi, O. (2016). A matter of time: small RNAs regulate the duration of epigenetic inheritance. *Trends Genet.* 33, 46–57. doi: 10.1016/j.tig.2016.11.001
- Huang, J. Y., Henao, O. L., Griffin, P. M., Vugia, D. J., Cronquist, A. B., Hurd, S., et al. (2016). Infection with pathogens transmitted commonly through food and the effect of increasing use of culture-independent diagnostic tests on surveillance–foodborne diseases active surveillance network, 10 U.S. Sites, 2012–2015. *Morb. Mortal. Wkly. Rep.* 65, 368–371. doi: 10.15585/mmwr.mm6514a2
- Jablonska, E., and Lamb, M. J. (2005). *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge, MA: MIT Press.
- Jackson, B. R., Tarr, C., Strain, E., Jackson, K. A., Conrad, A., Carleton, H., et al. (2016). Implementation of nationwide real-time whole-genome sequencing to

- enhance listeriosis outbreak detection and investigation. *Clin. Infect. Dis.* 63, 380–386. doi: 10.1093/cid/ciw242
- Janda, J. M., and Abbott, S. A. (2014). Culture-independent diagnostic testing: have we opened Pandora's box for good? *Diagn. Microbiol. Infect. Dis.* 80, 171–176. doi: 10.1016/j.diagmicrobio.2014.08.001
- Knowles, M., Stinson, S., Lambert, D., Carrillo, C., Koziol, A., Gauthier, M., et al. (2016). Genomic tools for customized recovery and detection of foodborne shiga toxicigenic *Escherichia coli*. *J. Food Prot.* 79, 2066–2077. doi: 10.4315/0362-028X.JFP-16-220
- Kwong, J., Mercoula, K., Tomita, T., Easton, M., Li, H. Y., Bulach, D. M., et al. (2016). Prospective whole-genome sequencing enhances national surveillance of *Listeria monocytogenes*. *J. Clin. Microbiol.* 54, 333–342. doi: 10.1128/JCM.02344-15
- Langley, G., Besser, J., Iwamoto, M., Lessa, F. C., Cronquist, A., Skoff, T. H., et al. (2015). Effect of culture-independent diagnostic tests on future emerging infections program surveillance. *Emerg. Infect. Dis.* 21, 1582–1588. doi: 10.3201/eid2109.150570
- Lindsey, R. L., Pouselee, H., Chen, J. C., Strockbine, N. A., and Carleton, H. A. (2016). Implementation of whole genome sequencing (WGS) for identification and characterization of shiga toxin-producing *Escherichia coli* (STEC) in the United States. *Front. Microbiol.* 7:766. doi: 10.3389/fmicb.2016.00766
- McKusick, V. A., and Ruddle, F. H. (1987). A new discipline, a new name, a new journal. *Genomics* 1, 1–2. doi: 10.1016/0888-7543(87)90098-X
- Murray, H. (2016). Connecting chromosome replication with cell growth in bacteria. *Curr. Opin. Microbiol.* 34, 13–17. doi: 10.1016/j.mib.2016.07.013
- Murray, H., and Koh, A. (2014). Multiple regulatory systems coordinate DNA replication with cell growth in *Bacillus subtilis*. *PLoS Genet.* 10:e1004731. doi: 10.1371/journal.pgen.1004731
- Pace, N. R. (2009). Mapping the tree of life: progress and prospects. *Microbiol. Mol. Biol. Rev.* 73, 565–576. doi: 10.1128/MMBR.00033-09
- Pallarès, I., Iglesias, V., and Ventura, S. (2015). The Rho termination factor of *Clostridium botulinum* contains a prion-like domain with a highly amyloidogenic core. *Front. Microbiol.* 6:1516. doi: 10.3389/fmicb.2015.01516
- Pettengill, J. B., Luo, Y., Davis, S., Chen, Y., Gonzalez-Escalona, N., Ottesen, A., et al. (2014). An evaluation of alternative methods for constructing phylogenies from whole genome sequence data: a case study with *Salmonella*. *PeerJ* 2:e620. doi: 10.7717/peerj.620
- Pinto, D., Santos, M. A., and Chambel, L. (2015). Thirty years of viable but nonculturable state research: unsolved molecular mechanisms. *Crit. Rev. Microbiol.* 41, 61–76. doi: 10.3109/1040841X.2013.794127
- Renesto, P., Crapoulet, N., Ogata, H., La Scola, B., Vestris, G., Claverie, J. M., et al. (2003). Genome-based design of a cell-free culture medium for *Tropheryma whipplei*. *Lancet* 362, 447–449. doi: 10.1016/S0140-6736(03)14071-8
- Reyes-Lamothe, R., Nicolas, E., and Sherratt, D. J. (2012). Chromosome replication and segregation in bacteria. *Annu. Rev. Genet.* 46, 121–143. doi: 10.1146/annurev-genet-110711-155421
- Ronholm, J., Nasheri, N., Petronella, N., and Pagotto, F. (2016). Navigating microbiological food safety in the era of whole-genome sequencing. *Clin. Microbiol. Rev.* 29, 837–857. doi: 10.1128/CMR.00056-16
- Sanchez-Vizcute, P., Orgaz, B., Aymerich, S., Le Coq, D., and Briandet, R. (2015). Pathogens protection against the action of disinfectants in multispecies biofilms. *Front. Microbiol.* 6:705. doi: 10.3389/fmicb.2015.00705
- Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., et al. (2004). Quality assessment of the human genome sequence. *Nature* 429, 365–368. doi: 10.1038/nature02390
- Todd, E. C. D., Greig, J. D., Bartleson, C. A., and Michaels, B. S. (2008). Outbreaks where food workers have been implicated in the spread of foodborne disease. Part 4. Infective doses and pathogen carriage. *J. Food Prot.* 71, 2339–2373. doi: 10.4315/0362-028X-71.11.2339
- USDA (2016). *Microbiology Laboratory Guidebook*. Available at: <http://www.fsis.usda.gov/wps/portal/fsis/topics/science/laboratories-and-procedures/guidebooks-and-methods/microbiology-laboratory-guidebook/microbiology-laboratory-guidebook> [accessed November 24, 2016].
- Van Acker, H., and Coenye, T. (2016). The role of efflux and physiological adaptation in biofilm tolerance and resistance. *J. Biol. Chem.* 291, 12565–12572. doi: 10.1074/jbc.R115.707257
- Verstraeten, N., Knapen, W., Fauvert, M., and Michiels, J. (2016). A historical perspective on bacterial persistence. *Methods Mol. Biol.* 1333, 3–13. doi: 10.1007/978-1-4939-2854-5_1
- Wang, Q., Holmes, N., Martinez, E., Howard, P., Hill-Cawthorne, G., and Sintchenko, V. (2015). It is not all about single nucleotide polymorphisms: comparison of mobile genetic elements and deletions in *Listeria monocytogenes* genomes links cases of hospital-acquired listeriosis to the environmental source. *J. Clin. Microbiol.* 53, 3492–3500. doi: 10.1128/JCM.00202-15
- Wang, Y., and Salazar, J. K. (2016). Culture-independent rapid detection methods for bacterial pathogens and toxins in food matrices. *Compr. Rev. Food Sci. Food Saf.* 15, 183–205. doi: 10.1111/1541-4337.12175
- Weedmark, K. A., Mabon, P., Hayden, K. L., Lambert, D., Van Domselaar, G., Austin, J. W., et al. (2015). *Clostridium botulinum* group II isolate phylogenomic profiling using whole-genome sequence data. *Appl. Environ. Microbiol.* 81, 5938–5948. doi: 10.1128/AEM.01155-15
- Whiteside, M. D., Laing, C. R., Manji, A., Kruczakiewicz, P., Taboada, E. N., and Gannon, V. P. (2016). SuperPhy: predictive genomics for the bacterial pathogen *Escherichia coli*. *BMC Microbiol.* 16:65. doi: 10.1186/s12866-016-0680-0
- Wilson, D. N. (2014). Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nat. Rev. Microbiol.* 12, 35–48. doi: 10.1038/nrmicro3155
- Wu, V. C. H. (2008). A review of microbial injury and recovery methods in food. *Food Microbiol.* 25, 735–744. doi: 10.1016/j.fm.2008.04.011
- Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The SILVA and “all-species living tree project (LTP)” taxonomic frameworks. *Nucleic Acids Res.* 42, D643–D648. doi: 10.1093/nar/gkt1209
- Yoshida, C., Kruczakiewicz, P., Laing, C. R., Lingohr, E. J., Gannon, V. P. J., Nash, J. H. E., et al. (2016). The *Salmonella* *in silico* typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS ONE* 11:e0147101. doi: 10.1371/journal.pone.0147101
- Zumla, A., Al-Tawfiq, J. A., Enne, V. I., Kidd, M., Drosten, C., Breuer, J., et al. (2014). Rapid point of care diagnostic tests for viral and bacterial respiratory tract infections—needs, advances, and future prospects. *Lancet Infect. Dis.* 14, 1123–1135. doi: 10.1016/S1473-0994(14)70827-8

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Gill. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance

Emma Griffiths¹, Damion Dooley², Morag Graham^{3,4}, Gary Van Domselaar^{3,4}, Fiona S. L. Brinkman¹ and William W. L. Hsiao^{2,5*}

¹ Department of Molecular Biology and Biochemistry, Simon Fraser University, Vancouver, BC, Canada, ² Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada, ³ National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada, ⁴ Department of Medical Microbiology and Infectious Diseases, Max Rady College of Medicine, University of Manitoba, Winnipeg, MB, Canada, ⁵ British Columbia Centre for Disease Control Public Health Laboratory, Vancouver, BC, Canada

OPEN ACCESS

Edited by:

Jennifer Ronholm,
McGill University, Canada

Reviewed by:

Roberto Spreafico,
Synthetic Genomics,
United States

Abasiofiok Mark Ibekwe,
Agricultural Research Service (USDA),
United States

***Correspondence:**

William W. L. Hsiao
william.hsiao@bccdc.ca

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 21 February 2017

Accepted: 29 May 2017

Published: 26 June 2017

Citation:

Griffiths E, Dooley D, Graham M, Van Domselaar G, Brinkman FSL and Hsiao WWL (2017) Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance.

Front. Microbiol. 8:1068.
doi: 10.3389/fmicb.2017.01068

Globalization of food networks increases opportunities for the spread of foodborne pathogens beyond borders and jurisdictions. High resolution whole-genome sequencing (WGS) subtyping of pathogens promises to vastly improve our ability to track and control foodborne disease, but to do so it must be combined with epidemiological, clinical, laboratory and other health care data (called “contextual data”) to be meaningfully interpreted for regulatory and health interventions, outbreak investigation, and risk assessment. However, current multi-jurisdictional pathogen surveillance and investigation efforts are complicated by time-consuming data re-entry, curation and integration of contextual information owing to a lack of interoperable standards and inconsistent reporting. A solution to these challenges is the use of ‘ontologies’ - hierarchies of well-defined and standardized vocabularies interconnected by logical relationships. Terms are specified by universal IDs enabling integration into highly regulated areas and multi-sector sharing (e.g., food and water microbiology with the veterinary sector). Institution-specific terms can be mapped to a given standard at different levels of granularity, maximizing comparability of contextual information according to jurisdictional policies. Fit-for-purpose ontologies provide contextual information with the auditability required for food safety laboratory accreditation. Our research efforts include the development of a Genomic Epidemiology Ontology (GenEpiO), and Food Ontology (FoodOn) that harmonize important laboratory, clinical and epidemiological data fields, as well as existing food resources. These efforts are supported by a global consortium of researchers and stakeholders worldwide. Since foodborne diseases do not respect international borders, uptake of such vocabularies will be crucial for multi-jurisdictional interpretation of WGS results and data sharing.

Keywords: genomic epidemiology, foodborne pathogen surveillance, outbreak investigations, ontology, contextual metadata

INTRODUCTION: THE IMPORTANCE OF METADATA AND CONTEXTUAL INFORMATION IN FOODBORNE SAFETY AND SURVEILLANCE

Foodborne pathogens impact global health and can cost economies millions of dollars in lost productivity (Flynn, 2014; Minor et al., 2015; World Health Organization, 2015). “Integrated surveillance” combines data from different stages of the farm-to-fork food continuum to provide multi-sector information for infectious disease surveillance, and represents the most comprehensive strategy to improve food safety (Zaidi et al., 2008; Ammon and Makela, 2010; Danan et al., 2011). Central to public health microbiology, food safety, and disease surveillance activities, is the comparison of genetic relatedness between isolates from human, food, and environmental samples. Whole genome sequencing (WGS) provides the highest resolution evidence for inferring phylogenetic relationships among foodborne pathogens (Ashton et al., 2016; Kanagarajah et al., 2017; Waldrum et al., 2017). However, genomic sequences can only be consistently interpreted for food safety and surveillance when the data are linked to standardized, fit-for-purpose contextual information suitable for use by data analysts, data consumers, and stakeholders (Lambert et al., 2017).

Contextual information in genomic epidemiology investigations includes critical knowledge about sequencing pipelines and sequence quality, sources of exposure and risk, clinical phenotypes, susceptible populations, geographical distribution and more. Reliable capture of parameters pertaining to sample provenance (specimen types and sources), sample processing (DNA extraction and sequencing library construction), quality control (sequence quality and contamination detection), data analysis (bioinformatic pipelines) are critical for reproducibility, comparability, and calibration of genomic results (Kircher et al., 2011; Paszkiewicz et al., 2014; Lynch et al., 2016). In addition to sequencing and bioinformatics parameters, laboratory test results characterizing antimicrobial resistance and virulence phenotypes often reveal important pathogen determinants that help to inform source and risk (World Health Organization, 2008; Clark et al., 2016; Glasser et al., 2016; Sharma et al., 2016; Day et al., 2017; Kanengoni et al., 2017; Tagini et al., 2017). Clinical information about the host, and epidemiological information about possible exposures (high-risk food types), are all useful to establish at-risk populations and hypothesize about likely sources of contamination (World Health Organization, 2008). This information is also used to establish the geographic distribution of pathogenic strains, as well as among populations, which is critical for determining transmission patterns (Moura et al., 2016; Njamkepo et al., 2016). Rich contextual information increases the utility of genomics data used for food safety surveillance, outbreak investigations, source attribution and risk assessments. Risk analysis in particular requires precise data on pathogen hazards in food to be systematically linked to epidemiological data, in order to make assessments, implement interventions and monitor outcomes (Lammerding and Fazil, 2000; Hoornstra

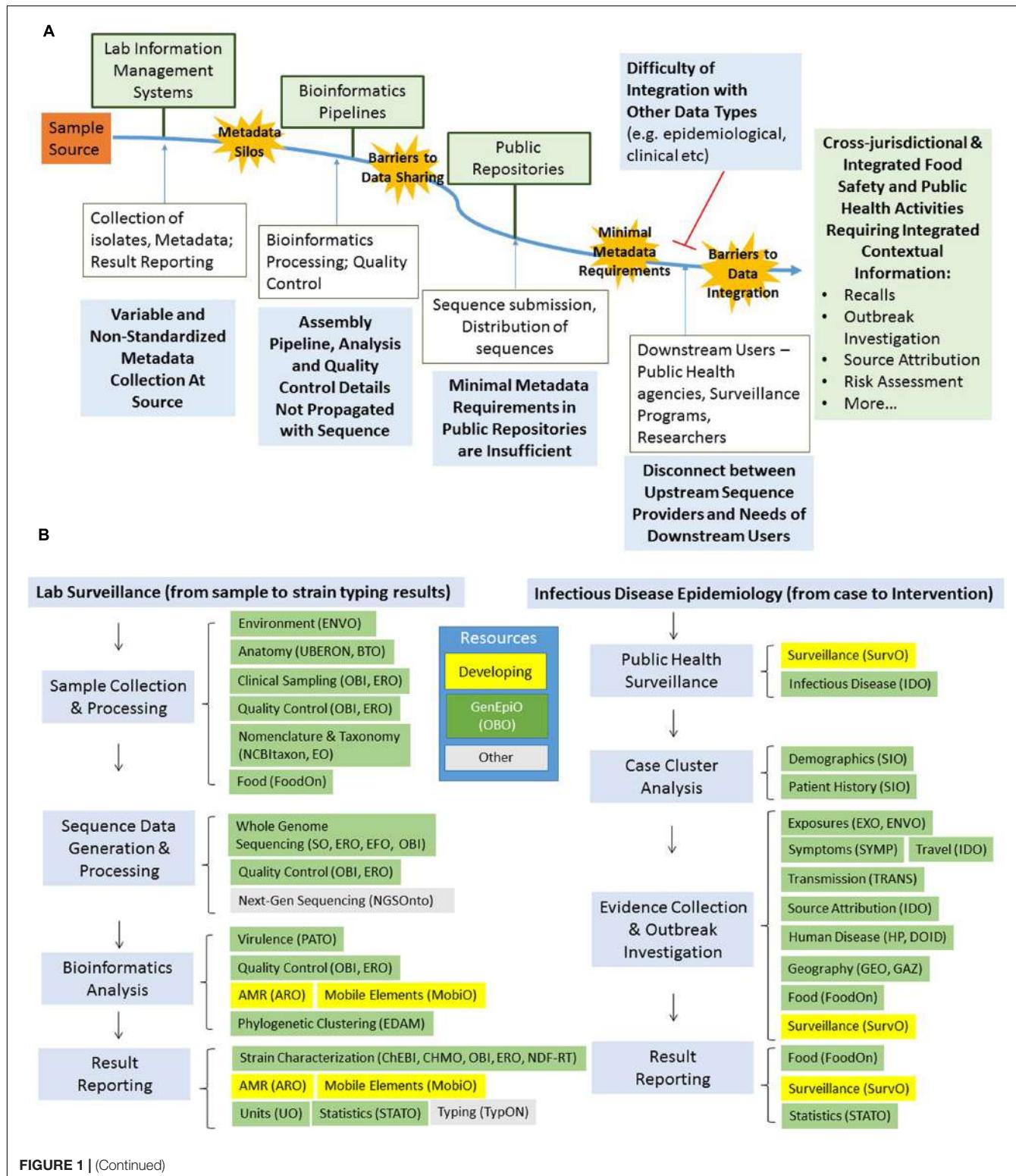
et al., 2001; Food and Agriculture Organization of the United Nations [FAO], 2005).

Unfortunately, resource-demands for the collection of such information, inconsistencies in descriptors, as well as other political and technical barriers have proven to complicate data sharing and integration between agencies. Wide adoption of contextual information best practices, as well as storage and sharing practices, would enable rapid, on-demand comparison of sequences from different sources and agencies, enhancing pathogen detection, inter-agency communication and responses. Here, we describe these various challenges and explain how informatics innovations such as ontologies can provide much needed solutions to streamline data interpretation and exchange for improved food safety and public health.

BARRIERS TO INTEGRATION AND SHARING OF WHOLE GENOME SEQUENCE DATA AND CONTEXTUAL INFORMATION

Despite a growing global commitment to the use and sharing of public health microbiology data, implementation at local, regional, national, and international levels has proven challenging with both political and technological barriers (van Panhuis et al., 2014). Fundamental structural barriers embedded in public health governance systems arise as the result of lack of trust (Pisani and AbouZahr, 2010; Fidler and Gostin, 2011; van Panhuis et al., 2014). Perceptions of risk to patient privacy and intellectual property, as well as the fear of misinterpretation and potential misuse of data are some of the biggest challenges to the sharing of sequence data and the exchange of contextual information (van Panhuis et al., 2014). Risk aversion practices prompt health agencies to implement blanket policies restricting data sharing, which result in incomplete metadata attached to sequences in public data repositories (van Panhuis et al., 2014).

Technological barriers for electronic data interchange exacerbate issues of political distrust (van Panhuis et al., 2014). Contextual data are mostly expressed as free text or agency-specific terminology. While reports and guidelines exist in an effort to suggest minimum contextual information that should be attached to genomic sequences, these fields are rarely incorporated into Lab Information Management Systems (LIMS) and epidemiology surveillance forms (Field et al., 2014; Grad and Lipsitch, 2014; Aziz et al., 2015; McMahon and Denaxas, 2016; Lambert et al., 2017). Through user interviews and needs assessments, we and others have found that information is then “siloed” in different hard drives, agencies, in restrictive data formats (paper or antiquated electronic formats), and is often collected for short-term purposes (van Panhuis et al., 2014). Owing to such inconsistency, recoding of the data is often needed for data sharing across institutions participating in multi-jurisdictional surveillance, impacting



response time. By relying on retrospective retrieval from different sources (as opposed to real-time collection), the quality and quantity of contextual information become eroded over

time. Flow of contextual information from source to end user, as well as barriers to collection and sharing are illustrated in Figure 1.

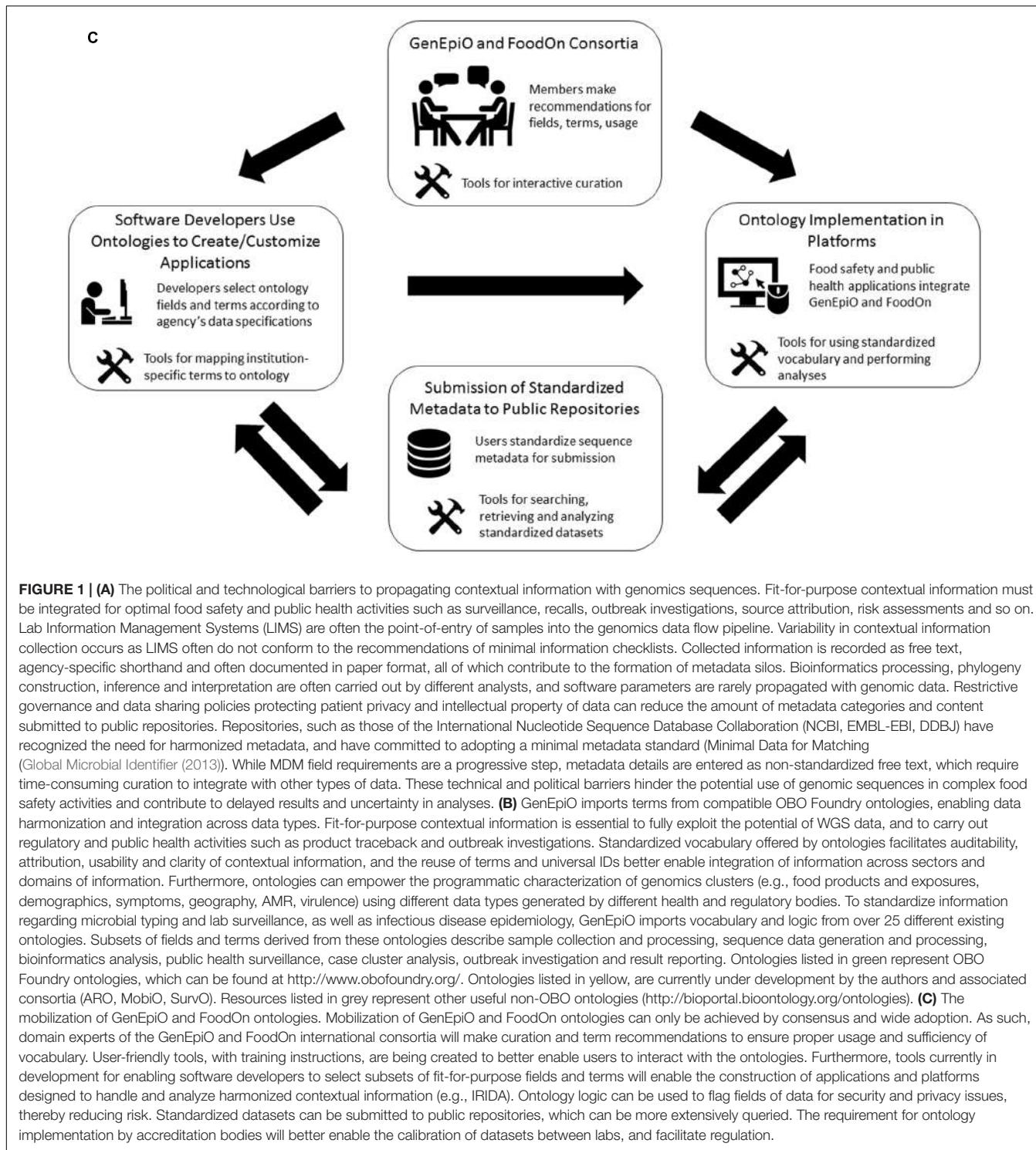


FIGURE 1 | (A) The political and technological barriers to propagating contextual information with genomics sequences. Fit-for-purpose contextual information must be integrated for optimal food safety and public health activities such as surveillance, recalls, outbreak investigations, source attribution, risk assessments and so on. Lab Information Management Systems (LIMS) are often the point-of-entry of samples into the genomics data flow pipeline. Variability in contextual information collection occurs as LIMS often do not conform to the recommendations of minimal information checklists. Collected information is recorded as free text, agency-specific shorthand and often documented in paper format, all of which contribute to the formation of metadata silos. Bioinformatics processing, phylogeny construction, inference and interpretation are often carried out by different analysts, and software parameters are rarely propagated with genomic data. Restrictive governance and data sharing policies protecting patient privacy and intellectual property of data can reduce the amount of metadata categories and content submitted to public repositories. Repositories, such as those of the International Nucleotide Sequence Database Collaboration (NCBI, EMBL-EBI, DDBJ) have recognized the need for harmonized metadata, and have committed to adopting a minimal metadata standard (Minimal Data for Matching (Global Microbial Identifier (2013)). While MDM field requirements are a progressive step, metadata details are entered as non-standardized free text, which require time-consuming curation to integrate with other types of data. These technical and political barriers hinder the potential use of genomic sequences in complex food safety activities and contribute to delayed results and uncertainty in analyses. **(B)** GenEpiO imports terms from compatible OBO Foundry ontologies, enabling data harmonization and integration across data types. Fit-for-purpose contextual information is essential to fully exploit the potential of WGS data, and to carry out regulatory and public health activities such as product traceback and outbreak investigations. Standardized vocabulary offered by ontologies facilitates auditability, attribution, usability and clarity of contextual information, and the reuse of terms and universal IDs better enable integration of information across sectors and domains of information. Furthermore, ontologies can empower the programmatic characterization of genomics clusters (e.g., food products and exposures, demographics, symptoms, geography, AMR, virulence) using different data types generated by different health and regulatory bodies. To standardize information regarding microbial typing and lab surveillance, as well as infectious disease epidemiology, GenEpiO imports vocabulary and logic from over 25 different existing ontologies. Subsets of fields and terms derived from these ontologies describe sample collection and processing, sequence data generation and processing, bioinformatics analysis, public health surveillance, case cluster analysis, outbreak investigation and result reporting. Ontologies listed in green represent OBO Foundry ontologies, which can be found at <http://www.obofoundry.org/>. Ontologies listed in yellow, are currently under development by the authors and associated consortia (ARO, MobiO, SurvO). Resources listed in grey represent other useful non-OBO ontologies (<http://bioportal.bioontology.org/ontologies>). **(C)** The mobilization of GenEpiO and FoodOn ontologies. Mobilization of GenEpiO and FoodOn ontologies can only be achieved by consensus and wide adoption. As such, domain experts of the GenEpiO and FoodOn international consortia will make curation and term recommendations to ensure proper usage and sufficiency of vocabulary. User-friendly tools, with training instructions, are being created to better enable users to interact with the ontologies. Furthermore, tools currently in development for enabling software developers to select subsets of fit-for-purpose fields and terms will enable the construction of applications and platforms designed to handle and analyze harmonized contextual information (e.g., IRIDA). Ontology logic can be used to flag fields of data for security and privacy issues, thereby reducing risk. Standardized datasets can be submitted to public repositories, which can be more extensively queried. The requirement for ontology implementation by accreditation bodies will better enable the calibration of datasets between labs, and facilitate regulation.

EXISTING RESOURCES FOR METADATA STANDARDIZATION AND FOOD SAFETY: FROM CHECKLISTS TO ONTOLOGIES

One of the biggest challenges to the standardization of metadata capture for food safety is the large number of

incompatible food classifications used worldwide. These food classifications range from lists of food types, descriptors of food production environments, codes of practice, guidelines, and other recommendations relating to foods, food production, and food safety. While these resources are certainly useful, they have been developed for specific uses, and fundamental

differences in their architecture limit interoperability. A selection of such food dictionaries can be found in **Table 1**. For example, analyses of foodborne outbreak data for source attribution requires the categorization of reported food vehicle. Variation in the way aetiological agents and foods are defined and categorized, even within a single country or jurisdiction, has been shown to impede direct comparison of food attribution across countries within similar time periods (Greig and Ravel, 2009). While up-to-date food safety best practices prescribe data collection systems to be sufficiently precise in order to minimize uncertainty, in reality, inconsistencies in descriptors pertaining to the host, pathogen, environment, and the underlying attributes of potentially contaminated foods, all contribute to uncertainty in data analyses and delay in public health action (Greig and Ravel, 2009).

In designing an approach to capture standardized metadata, it is critical to define what information about a sample is most informative for its intended use. This process is best achieved via engagement of a variety of end users - in this case food regulators, epidemiologists, lab analysts, bioinformaticians, at local, regional, national and international levels. Minimum Information (MI) checklists represent the sum of all essential data fields recommended by community experts and users, with controlled vocabularies used as 'allowed values' (Field and Sansone, 2006). A well-known genomic metadata standard is the MIxS checklist, a minimal metadata standard checklist developed by the Genomic Standards Consortium (GSC) for reporting information about any nucleotide sequence (Yilmaz et al., 2011). Similarly, the National Institute of Allergy and Infectious Diseases Genome Sequencing Center and Bioinformatics Resource Center (GSCID/BRC) Project and Sample Application Standard specifically addresses metadata types that should be attached to human pathogen genomic sequences (Dugan et al., 2014). Additionally, the Minimum Information about a Phylogenetic Analysis (MIAPA) represents a community-wide effort to develop minimal reporting standards for phylogenetic analyses (Leebens-Mack et al., 2006). These checklists contain a wide variety of descriptive fields; however, they currently lack standardized values to enter in the fields.

A more comprehensive mechanism for making metadata searchable and actionable, is through the use of 'ontologies' (Bodenreider and Stevens, 2006; Brinkman et al., 2010). Ontologies are hierarchies of well-defined and standardized vocabulary interconnected by logical relationships (Bodenreider and Stevens, 2006). These logical interconnections provide a layer of intelligence to query engines, making ontologies much more powerful than simple flat lists of terms. Terms and their definitions, are specified by universal IDs (Universal Resource Identifiers), which associate descriptors with particular usages and disambiguate meaning (Bodenreider and Stevens, 2006). Ontologies also incorporate synonyms of terms in the definitions and identifiers (IDs) e.g., biscuits (United Kingdom) and cookies (North America), enabling institutions to use their preferred terminology while simultaneously mapping terms to an ontology standard. The hierarchical structure enables comparison of entities at different levels of granularity (e.g., leafy

greens and spinach), which represents an important feature for evolving food safety investigations in which the hypothesized food vehicle is a moving target. Mapping to an ontology-based standard and reuse of universal IDs makes software implementing the ontology framework interoperable, enabling faster and more efficient data exchange (Arp et al., 2015). The reuse of terms and their IDs enables integration of different data types across domains (epidemiology, food, disease, agriculture, antimicrobial resistance, etc) and between agencies (Ferreira et al., 2013). Computer and human readable (in different natural languages), ontology hierarchies allow stakeholders to share data according to the level of granularity permitted by jurisdictional policies, and fields of information with legal or privacy issues can be flagged using ontology relations to increase security. Furthermore, fit-for-purpose ontologies provide contextual information with the auditability required for food safety and public health laboratory accreditation (Evans, 2015). Principles of good practice in ontology development have been put into practice within the framework of the Open Biomedical Ontologies consortium through its OBO Foundry initiative, which emphasizes collaborative development, interoperability and usability (Smith et al., 2007). Descriptors of genomic epidemiological processes have already been captured in a number of existing ontologies. Some examples include the Sequence Ontology (SO) (Eilbeck et al., 2005), the EDAM Bioinformatics Ontology (EDAM) (Ison et al., 2013), and DOID (Schriml et al., 2012), which describe sequences, genome assembly, and human disease. The Exposure, Epidemiology, Environment, Symptoms, and Transmission Ontologies (EXO, EPO, ENVO, SYMP, TRANS) describe types of exposures, facets of epidemiology, natural and built environments, clinical signs and symptoms, and modes of transmission (Mattingly et al., 2012; Pesquita et al., 2014; Buttigieg et al., 2016). Ontologies and other resources useful for genomic epidemiology are listed in **Table 1**.

Currently, no resource(s) integrate all the necessary components of a genomic epidemiology investigation. As such, our research efforts have focused on the development of a Genomic Epidemiology Ontology (GenEpiO), based on public health stakeholder interviews and the harmonization of important laboratory, clinical and epidemiological data fields, in collaboration with a consortium of researchers and end users. We are also actively developing, in collaboration with members of the international GenEpiO consortium, a Farm-to-Fork food ontology (FoodOn) aiming to harmonize existing food resources and describe food entities from point(s) of production/collection, through processing, distribution and consumption.

GenEpiO AND FoodOn: NEW DEVELOPMENTS IN FOOD SAFETY SEMANTICS

The Genomic Epidemiology Ontology (GenEpiO) is an ontology resource being developed according to the principles of the OBO Foundry, led by a partnership of Canadian scientists representing academic, provincial and federal public health interests. The objective of GenEpiO is to enable integration and

TABLE 1 | A selection of ontology and Minimum Information (MI) checklists for the standardization of genomics metadata and epidemiological, clinical, and laboratory contextual information.

Resource	Description	URL
Codex Alimentarius	<ul style="list-style-type: none"> Internationally recognized standards, codes of practice, guidelines Recommendations relating to foods, food production, and food safety Commissioned by the United Nations Food and Agriculture Organization 	http://www.fao.org/fao-who-codexalimentarius/codex-home/en/
Langual	<ul style="list-style-type: none"> Created by US FDA's Centre for Food Safety and Applied Nutrition 14 main facets, or hierarchies of descriptive terms (35,000 foods) Available in many languages. 	http://www.langual.org/
Food Ex2	<ul style="list-style-type: none"> Created by the European Food Safety Authority (EFSA) Food classification designed to facilitate food exposure assessment 	https://www.efsa.europa.eu/en/data/data-standardisation
USDA National Nutrient Database for Standard Reference	<ul style="list-style-type: none"> Food dictionary containing over 9000 foods Each item lists nutrient values and weights per portion 	https://ndb.nal.usda.gov/ndb/foods
Compendium of Analytical Methods	<ul style="list-style-type: none"> Created by Health Canada Food list containing several hundred items organized by food category Designed to foster compliance of the food industry with standards and guidelines relative to microbiological and extraneous material in foods 	http://www.hc-sc.gc.ca/fn-an/res-rech/analy-meth/microbio/volume1-eng.php
Food Commodity Classification Scheme	<ul style="list-style-type: none"> Created by the US Center for Disease Control Designed for source attribution studies 	http://www.ncbi.nlm.nih.gov/pubmed/19968563
The Agriculture Ontology (AgrO)	<ul style="list-style-type: none"> The ontology of agronomic practices, agronomic techniques, and agronomic variables used in agronomic experiments 	http://www.obofoundry.org/ontology/agro.html
Antimicrobial Resistance Ontology (ARO)	<ul style="list-style-type: none"> Ontology of antibiotics, resistance genes, and associated phenotypes 	https://card.mcmaster.ca/
Basic Formal Ontology (BFO)	<ul style="list-style-type: none"> Upper level ontology designed to support information retrieval, analysis and integration in scientific, and other domains 	http://www.obofoundry.org/ontology/bfo.html
BRENDA Tissue Ontology (BTO)	<ul style="list-style-type: none"> Structured controlled vocabulary for the source of an enzyme comprising tissues, cell lines, cell types, and cell cultures 	http://www.obofoundry.org/ontology/bto.html
Chemical Entities of Biological Interest Ontology (ChEBI)	<ul style="list-style-type: none"> Structured classification of molecular entities of biological interest focusing on 'small' chemical compounds 	http://www.obofoundry.org/ontology/chebi.html
Cell Ontology (CL)	<ul style="list-style-type: none"> Structured controlled vocabulary for cell types in animals 	http://www.obofoundry.org/ontology/cl.html
Human Disease Ontology (DOID)	<ul style="list-style-type: none"> Ontology for describing the classification of human diseases organized by etiology 	http://www.obofoundry.org/ontology/doid.html
EMBRACE Data and Methods Ontology (EDAM)	<ul style="list-style-type: none"> Ontology of common bioinformatics operations, topics, types of data including identifiers, and formats 	http://www.ontobee.org/ontology/EDAM
Environment Ontology (ENVO)	<ul style="list-style-type: none"> Contained descriptors of a range of food products and food production environments Limited in scope, based on user suggestions 	http://www.obofoundry.org/ontology/envo.html
Epidemiology (EPO)	<ul style="list-style-type: none"> Ontology designed to support the semantic annotation of epidemiology resources 	http://www.obofoundry.org/ontology/epo.html
Exposure (EXO)	<ul style="list-style-type: none"> Vocabularies for describing exposure data to inform understanding of environmental health 	http://www.obofoundry.org/ontology/exo.html

(Continued)

TABLE 1 | Continued

Resource	Description	URL
Foundational Model of Anatomy (FMA)	<ul style="list-style-type: none"> Ontology representing phenotypic structures of the human body 	http://www.obofoundry.org/ontology/fma.html
FooDB Ontology (FoodO)	<ul style="list-style-type: none"> Designed to represent the FooDB database describing food items and chemical composition (additives, ingredients, etc) 	http://aber-owl.net/ontology/FOODO
Food Ontology (FoodOn)	<ul style="list-style-type: none"> Farm-to-Fork descriptors of food entities and food production environments from point of production through processing, distribution and consumption Created by the FoodOn Consortium 	http://www.obofoundry.org/ontology/foodon.html http://foodontology.github.io/foodon/
Genomic Epidemiology Ontology (GenEpiO)	<ul style="list-style-type: none"> Controlled vocabulary for infectious disease surveillance and outbreak investigations implementing whole genome sequencing Ongoing development via the International GenEpiO Consortium 	http://www.genepio.org http://www.obofoundry.org/ontology/genepio.html
Infectious Disease Ontology (IDO)	<ul style="list-style-type: none"> Ontology describing entities relevant to both biomedical and clinical aspects of most infectious diseases 	https://bioportal.bioontology.org/ontologies/IDO
Next-Generation Sequencing Ontology (NGSOnto)	<ul style="list-style-type: none"> Structured vocabulary to capture the workflow of all the processes involved in a Next Generation Sequencing project 	https://bioportal.bioontology.org/ontologies/NGSONTO
Ontology for Biomedical Investigations (OBI)	<ul style="list-style-type: none"> Ontology for the description of life-science and clinical investigations 	http://www.obofoundry.org/ontology/obi.html
Phenotypic Quality Ontology (PATO)	<ul style="list-style-type: none"> Ontology of biomedical phenotypic qualities (properties, attributes or characteristics) 	http://www.obofoundry.org/ontology/pato.html
Relation Ontology (RO)	<ul style="list-style-type: none"> Biology-specific relations to connect entities and classes Intended for standardization across OBO Foundry Library of ontologies 	http://www.obofoundry.org/ontology/ro.html
The Sustainable Development Goals Interface Ontology (SDGIO)	<ul style="list-style-type: none"> The Sustainable Development Goals Interface Ontology of United Nation Environmental Program 	https://github.com/SDG-InterfaceOntology/sdgio
Sequence Ontology (SO)	<ul style="list-style-type: none"> Structured controlled vocabulary for sequence annotation, for the exchange of annotation data and for the description of sequence objects in databases 	http://www.obofoundry.org/ontology/so.html
Systematized Nomenclature of Medicine (SNOMED)	<ul style="list-style-type: none"> Represents clinical phrases captured by the clinician Created by The International Health Terminology Standards Development Organisation (IHTSDO) 	http://www.ihtsdo.org/snomed-ct
Clinical Signs and Symptoms Ontology (SYMP)	<ul style="list-style-type: none"> Ontology to provide robust means to disambiguate, capture and document clinical signs, and symptoms 	http://www.obofoundry.org/ontology/symp.html
Pathogen Transmission Ontology (TRANS)	<ul style="list-style-type: none"> Ontology for describing transmission methods of human disease pathogens, from one host, reservoir, or source to another host 	http://www.obofoundry.org/ontology/trans.html
Microbial Typing Ontology (TypOn)	<ul style="list-style-type: none"> Structured vocabulary to describe microbial typing methods for the identification of bacterial isolates and their classification 	https://bioportal.bioontology.org/ontologies/TYPON
Multi-Species Anatomy Ontology (UBERON)	<ul style="list-style-type: none"> Integrated cross-species anatomy ontology covering animals and bridging multiple species-specific ontologies 	http://www.obofoundry.org/ontology/uberon.html
MlxS	<ul style="list-style-type: none"> A minimal metadata standard checklist developed by the Genomic Standards Consortium (GSC) for reporting information about any (x) nucleotide sequence 	Yilmaz et al., 2011

(Continued)

TABLE 1 | Continued

Resource	Description	URL
Project and Sample Application Standard	<ul style="list-style-type: none"> Created by the National Institute of Allergy and Infectious Disease Genome Sequencing Center and Bioinformatics Resource Center (GSCID/BRC) Specifically addresses metadata types that should be attached to human pathogen genomic sequences 	Dugan et al., 2014
Minimum Information about a Phylogenetic Analysis (MIAPA)	<ul style="list-style-type: none"> Community-wide effort to develop minimal reporting standards for phylogenetic analyses 	Leebens-Mack et al., 2006
STROME-ID guidelines	<ul style="list-style-type: none"> "Strengthening the reporting of molecular epidemiology for infectious diseases" Standards for reporting molecular epidemiology results including measures of genetic diversity, laboratory methods, sample collection, etc 	Field et al., 2014
The Global Alliance for Genomics and Health (GA4GH)	<ul style="list-style-type: none"> Aim to create a common, harmonized framework to enable secure sharing of genomic and clinical data 	http://genomicsandhealth.org/
The Global Microbial Identifier (GMI)	<ul style="list-style-type: none"> Platform for storing whole genome sequencing (WGS) data of microorganisms to detect outbreaks and emerging pathogens 	http://www.globalmicrobialidentifier.org/
The United Nations Environment Programme (UNEP)	<ul style="list-style-type: none"> Leading global environmental authority Promotes the coherent implementation of actions for sustainable development (Sustainable Development Goals) 	http://web.unep.org/
United Nations Environment Live	<ul style="list-style-type: none"> Interactive platform for environmental assessments and peer review of the SDGIO 	https://uneplive.unep.org/sdgs

propagation of all necessary contextual information required to interpret microbial pathogen genomics data, from the point-of-sample-intake, through sequencing, to end use (e.g., during a foodborne outbreak investigation). The GenEpiO hierarchy was constructed based on the Basic Formal Ontology (BFO) and Relation Ontology (RO) of the OBO Foundry, which delineate how *things* should be organized into higher level *classes*, and how *things* and *classes* should relate to one another (Smith et al., 2005; Arp et al., 2015). This architecture improves compatibility with other OBO biomedical ontologies, enriching vocabulary and data linkages, and facilitating the reuse of terminology and the integration of information across health and food safety domains (agriculture, veterinary care, environment, food production). The considerable consensus achieved by the OBO Foundry has paved the way for harmonization of complex content in a way that is unavailable with other disparate ontologies. GenEpiO terms are mapped to community standards and over 25 existing ontologies to ensure the accuracy of meaning and to facilitate interoperability (**Figure 1B**). GenEpiO also includes data models comprising disease/agency/reporting or analytical system/surveillance network-specific fields, which can be used to represent genomic epidemiology workflows, processes, disease progression and decision-making. GenEpiO currently contains over 2000 key fields and terms to harmonize sample metadata, lab analytics, wet lab and bioinformatics processes, quality control, clinical information as well as exposures and epidemiological data. As such, we anticipate that GenEpiO will better enable the calibration and validation of genomics for clinical and regulatory use. Controlled vocabulary and

relationship logic are encoded in the Web Ontology Language, OWL. OWL files are publicly available, and can be implemented in different software applications (**Table 1**). The GenEpiO ontology is currently being implemented within the Integrated Rapid Infectious Disease Analysis (IRIDA) platform¹, an open source, secure web-based, end-to-end platform for infectious disease genomic epidemiology, spearheaded in Canada. Within IRIDA, GenEpiO is being used to generate NCBI BioSample-compliant submission-ready genome metadata files, and to create different Line List visualization tools for epidemiological investigations. The next phase of development will involve the complete integration of GenEpiO to enhance the platform's analytical power.

FoodOn encompasses materials in natural ecosystems, as well as human-centric food items, food production environments and handling of food (Griffiths et al., 2016). We aim to develop semantics for food safety, food security, the agricultural and animal husbandry practices linked to food production, culinary, nutritional and chemical ingredients and processes. As such, FoodOn architecture is similarly based on BFO and RO schema, as well as the facet-based LanguaL (*Langua aLimentaria*, or language of food) classification system of the US Food and Drug Administration (US FDA) (Ireland and Møller, 2010). Facets include Food Products, which can be linked to Food Sources, Cooking and Preservation Methods, Consumer Groups, Cultural Origins, Taxonomy and more. Thousands of individual food products have already been indexed according to the

¹www.irida.ca

LanguaL system, and are publicly available in a separate FoodOn import file (**Table 1**). The scope of FoodOn is ambitious and will require input and long-term development by multiple domain experts. Further details regarding GenEpiO and FoodOn design and content will be discussed elsewhere (manuscripts in preparation).

In order to ensure utility, accuracy and usability, user engagement is a top priority for GenEpiO and FoodOn development. Feedback from engagement efforts has indicated that user-friendly tools for curation of terms, implementation, and mapping between interfaces and agencies, would serve to mobilize these technologies. To that effect, we are currently developing software applications for ontology mapping and curation. Additionally, both ontologies can be searched using various widely used portals such as the EBI Ontology Look-up Service, Ontobee, and NCBO BioPortal (**Table 1**). As harmonization of the both GenEpiO and FoodOn ontologies can only be achieved by consensus and wide adoption, involving open source and open access initiatives, we have catalyzed the formation of international consortia to build partnerships and solicit contributions from domain experts. The GenEpiO consortium membership comprises over 70 participants from 15 countries, with leadership, technical and editorial working groups. The interaction of the consortia, tools, applications, ontologies, users and repositories will be important for soliciting term contributions, as well as integrating regional- and sector-specific vocabulary, and evolving strategies for international uptake (**Figure 1C**).

BROADER CONTEXT OF FOOD GENOMICS METADATA AND ONTOLOGIES

Several frameworks for integrating genomics and other data currently exist for tackling the real-world problems of emerging diseases, environmental degradation, world hunger, and sustainability. Each of these global partnerships seeks to streamline the flow of genomics knowledge and its application for solving global challenges. The Global Alliance for Genomics and Health (GA4GH) and The Global Microbial Identifier (GMI) work to establish common frameworks and transdisciplinary networks to better monitor and control emerging public health threats (Knoppers, 2014; Wielinga et al., 2017). The Environmental Working Group of the United Nations (UNEP) have developed Sustainable Development Goals addressing climate change, renewable energy, food, health and water provision requiring the coordinated global monitoring (United Nations, 2016). Each of these efforts involves highly negotiated language representing different disciplines and policies, which can be harmonized into a coherent system through the use of ontologies. GA4GH and UNEP currently implement OBO Foundry ontologies that have been integrated into GenEpiO (e.g., ENVO, UBERON, ChEBI). GenEpiO integrates the Minimal Data for Matching standards for matching pathogen isolates prescribed by the GMI consortium (Global Microbial

Identifier, 2013), and GenEpiO and FoodOn standards are being considered for an upcoming ISO (International Organization for Standards) guideline on the use of WGS for Food Safety. The standardized food and food environment descriptors being developed in FoodOn can fill a critical gap in community standards required to integrate food related data in each of these efforts. Global initiatives and associated ontologies can be found in **Table 1**. Public health and genomics descriptors found in GenEpiO, combined with existing compatible ontologies for describing different environments (ENVO), agriculture (AgrO), and sustainable development (SDGIO), will greatly enable the integration of knowledge required to accomplish global health, equity and sustainability goals (**Table 1**).

CONCLUSION

Platforms implementing ontologies such as GenEpiO and FoodOn will be the work-engines ensuring the integration and reusability of genomics data from the collection of samples, through consumption by various end users. With the international nature of food distribution and food safety concerns, the most effective semantic resources must be open source, interoperable and collaboratively developed in order to best represent the needs of the international community. Global networks navigating the political challenges inherent in such community efforts will be crucial for the success of genomics as the new currency of food and waterborne pathogen typing. While no “one-size-fits-all” data dictionary for genomic epidemiology currently exists, harmonization of different vocabularies can be achieved through the use of ontologies and the flexibility they provide. With growing support of community-based development efforts, this foundational work can facilitate intra- and international data exchange, resulting in improved food safety and health outcomes globally, as well as promoting innovation and discovery.

AUTHOR CONTRIBUTIONS

EG wrote the manuscript. EG and DD developed software, concepts and resources. MG and GVD contributed input, use cases and testing material for resource development. WH and FB conceived the project and supervised this work. DD, MG, GVD, FB, and WH provided feedback on the manuscript.

FUNDING

This work was funded by Genome Canada Bioinformatics and Computational Biology (BCB) 2012 Grant #172PHM with co-funding from Genome BC and the federal Genomics Research and Development Initiative (GRDI) interdepartmental Food and Water Safety project. FoodON is funded by Genome Canada BCB 2015 Grant #254EPI, with some additional support from AllerGen NCE, Inc., of the Government of Canada’s Networks of Centres of Excellence (NCE) program.

ACKNOWLEDGMENTS

The authors would like to thank the GenEpiO Consortium for their contributions and support, as well as Pier Luigi Buttigieg,

REFERENCES

- Ammon, A., and Makela, P. (2010). Integrated data collection on zoonoses in the European Union, from animals to humans, and the analyses of the data. *Int. J. Food Microbiol.* 139(Suppl. 1), S43–S47. doi: 10.1016/j.ijfoodmicro.2010.03.002
- Arp, R., Smith, B., and Spear, A. D. (2015). *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: The MIT Press.
- Ashton, P. M., Nair, S., Peters, T. M., Bale, J. A., Powell, D. G., Painset, A., et al. (2016). Identification of *Salmonella* for public health surveillance using whole genome sequencing. *PeerJ* 4:e1752. doi: 10.7717/peerj.1752
- Aziz, N., Zhao, Q., Bry, L., Driscoll, D. K., Funke, B., Gibson, J. S., et al. (2015). College of american pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch. Pathol. Lab. Med.* 139, 481–493. doi: 10.3760/cma.j.issn.0529-5815.2017.02.004
- Bodenreider, O., and Stevens, R. (2006). Bio-ontologies: current trends and future directions. *Brief. Bioinform.* 7, 256–274. doi: 10.1093/bib/bbl027
- Brinkman, R. R., Courtot, M., Derom, D., Fostel, J. M., He, Y., Lord, P., et al. (2010). Modeling biomedical experimental processes with OBI. *J. Biomed. Semant.* 1(Suppl. 1), S7. doi: 10.1186/2041-1480-1-S1-S7
- Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., and Mungall, C. J. (2016). The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semant.* 7:57. doi: 10.1186/s13326-016-0097-6
- Clark, C. G., Berry, C., Walker, M., Petkau, A., Barker, D. O. R., Guan, C., et al. (2016). Genomic insights from whole genome sequencing of four clonal outbreak *Campylobacter jejuni* assessed within the global *C. jejuni* population. *BMC Genomics* 17:990. doi: 10.1186/s12864-016-3340-8
- Danan, C., Baroukh, T., Moury, F., Jourdan-Da Silva, N., Brisabois, A., and Le Strat, Y. (2011). Automated early warning system for the surveillance of *Salmonella* isolated in the agro-food chain in France. *Epidemiol. Infect.* 139, 736–741. doi: 10.1017/S0950268810001469
- Day, M., Doumith, M., Jenkins, C., Dallman, T. J., Hopkins, K. L., Elson, R., et al. (2017). Antimicrobial resistance in Shiga toxin-producing *Escherichia coli* serogroups O157 and O26 isolated from human cases of diarrhoeal disease in England, 2015. *J. Antimicrob. Chemother.* 72, 145–152. doi: 10.1093/jac/dkw371
- Dugan, V. G., Emrich, S. J., Giraldo-Calderón, G. I., Harb, O. S., Newman, R. M., Pickett, B. E., et al. (2014). Standardized metadata for human pathogen/vector genomic sequences. *PLoS ONE* 9:e99979. doi: 10.1371/journal.pone.0099979
- Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., et al. (2005). The Sequence ontology: a tool for the unification of genome annotations. *Genome Biol.* 6:R44. doi: 10.1186/gb-2005-6-5-r44
- Evans, P. (2015). “International standards development for use of whole genome sequencing in food microbiology,” in *Proceedings of the InFORM Meeting*, Phoenix, AZ.
- Ferreira, J. D., Paolotti, D., Couto, F. M., and Silva, M. J. (2013). On the usefulness of ontologies in epidemiology research and practice. *J. Epidemiol. Commun. Health* 67, 385–388. doi: 10.1136/jech-2012-201142
- Fidler, D. P., and Gostin, L. O. (2011). The WHO pandemic influenza preparedness framework: a milestone in global governance for health. *JAMA* 306, 200–201. doi: 10.1001/jama.2011.960
- Field, D., and Sansone, S.-A. (2006). A special issue on data standards. *OMICS J. Integr. Biol.* 10, 84–93. doi: 10.1089/omi.2006.10.84
- Field, N., Cohen, T., Struelens, M. J., Palm, D., Cookson, B., Glynn, J. R., et al. (2014). Strengthening the reporting of molecular epidemiology for infectious diseases (STROME-ID): an extension of the STROBE statement. *Lancet Infect. Dis.* 14, 341–352. doi: 10.1016/S1473-3099(13)70324-4
- Flynn, D. (2014). USDA: U.S. foodborne illnesses cost more than \$15.6 billion annually. *Food Saf. News*. Available at: <http://www.foodsafetynews.com/2014/10/foodborne-illnesses-cost-usa-15-6-billion-annually/>
- Robert Hoehndorf, Matthew Lange and Chris Mungall of the FoodOn Consortium, and Jane Ireland and Anders Møller of The Danish Food Informatics (DFI) group, for their ongoing development efforts.
- Food and Agriculture Organization of the United Nations [FAO] (2005). *Food Safety Risk Analysis - An Overview and Framework Manual*. Available at: https://www.fao.org/sonota/foodsafety_riskanalysis.pdf
- Glasset, B., Herbin, S., Guillier, L., Cadel-Six, S., Vignaud, M.-L., Grout, J., et al. (2016). *Bacillus cereus*-induced food-borne outbreaks in France, 2007 to 2014: epidemiology and genetic characterisation. *Euro. Surveill.* 21:30413. doi: 10.2807/1560-7917.ES.2016.21.48.30413
- Global Microbial Identifier (2013). *6th Annual Meeting on Global Microbial Identifier*. Sacramento, CA: Global Microbial Identifier. Available at: <http://www.globalmicrobialidentifier.org/news-and-events/previous-meetings/6th-meeting-on-gmi>
- Grad, Y. H., and Lipsitch, M. (2014). Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol.* 15:538. doi: 10.1186/s13059-014-0538-4
- Greig, J. D., and Ravel, A. (2009). Analysis of foodborne outbreak data reported internationally for source attribution. *Int. J. Food Microbiol.* 130, 77–87. doi: 10.1016/j.ijfoodmicro.2008.12.031
- Griffiths, E., Dooley, D., Buttigieg, P. L., Hoehndorf, R., Brinkman, F., and Hsiao, W. (2016). “FoodOn: a global farm-to-fork food ontology,” in *Proceedings of the ICBO Conference*, Corvallis, OR.
- Hoornstra, E., Northolt, M. D., Notermans, S., and Barendsz, A. W. (2001). The use of quantitative risk assessment in HACCP. *Food Control* 12, 229–234. doi: 10.1016/j.ijfoodmicro.2001.03.032
- Ireland, J. D., and Møller, A. (2010). LanguaL food description: a learning process. *Eur. J. Clin. Nutr.* 64, S44–S48. doi: 10.1038/ejcn.2010.209
- Ison, J., Kalas, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., et al. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics* 29, 1325–1332. doi: 10.1093/bioinformatics/btt113
- Kanagarajah, S., Waldrum, A., Dolan, G., Jenkins, C., Ashton, P. M., Carrion Martin, A. I., et al. (2017). Whole genome sequencing reveals an outbreak of *Salmonella* Enteritidis associated with reptile feeder mice in the United Kingdom, 2012–2015. *Food Microbiol.* (in press).
- Kanengoni, A. T., Thomas, R., Gelaw, A. K., and Madoroba, E. (2017). Epidemiology and characterization of *Escherichia coli* outbreak on a pig farm in South Africa. *FEMS Microbiol. Lett.* 364:fnx010. doi: 10.1093/femsle/fnx010
- Kircher, M., Heyn, P., and Kelso, J. (2011). Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics* 12:382. doi: 10.1186/1471-2164-12-382
- Knoppers, B. M. (2014). Framework for responsible sharing of genomic and health-related data. *HUGO J.* 8:3. doi: 10.1186/s11568-014-0003-1
- Lambert, D., Pightling, A., Griffiths, E., Van Domselaar, G., Evans, P., Berthelet, S., et al. (2017). Baseline practices for the application of genomic data supporting regulatory food safety. *J. AOAC Int.* 100, 721–731. doi: 10.5740/jaoacint.16-0269
- Lammerding, A. M., and Fazil, A. (2000). Hazard identification and exposure assessment for microbial food safety risk assessment. *Int. J. Food Microbiol.* 58, 147–157. doi: 10.1016/S0168-1605(00)00269-5
- Leebens-Mack, J., Vision, T., Brenner, E., Bowers, J. E., Cannon, S., Clement, M. J., et al. (2006). Taking the first steps towards a standard for reporting on phylogenies: minimum information about a phylogenetic analysis (MIAPA). *Oncics J. Integr. Biol.* 10, 231–237. doi: 10.1089/omi.2006.10.231
- Lynch, T., Petkau, A., Knox, N., Graham, M., and Domselaar, G. V. (2016). A primer on infectious disease bacterial genomics. *Clin. Microbiol. Rev.* 29, 881–913. doi: 10.1128/CMR.00001-16
- Mattingly, C. J., McKone, T. E., Callahan, M. A., Blake, J. A., and Hubal, E. A. C. (2012). Providing the missing link: the exposure science ontology ExO. *Environ. Sci. Technol.* 46, 3046–3053. doi: 10.1021/es2033857
- McMahon, C., and Denaxas, S. (2016). A novel framework for assessing metadata quality in epidemiological and public health research settings. *AMIA Summits Transl. Sci. Proc.* 2016, 199–208.

- Minor, T., Lasher, A., Klontz, K., Brown, B., Nardinelli, C., and Zorn, D. (2015). The per case and total annual costs of foodborne illness in the United States. *Risk Anal.* 35, 1125–1139. doi: 10.1111/risa.12316
- Moura, A., Criscuolo, A., Pouseele, H., Maury, M. M., Leclercq, A., Tarr, C., et al. (2016). Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nat. Microbiol.* 2:16185. doi: 10.1038/nmicrobiol.2016.185
- Njamkepo, E., Fawal, N., Tran-Dien, A., Hawkey, J., Strockbine, N., Jenkins, C., et al. (2016). Global phylogeography and evolutionary history of *Shigella dysenteriae* type 1. *Nat. Microbiol.* 1:16027. doi: 10.1038/nmicrobiol.2016.27
- Paszkiewicz, K. H., Farbos, A., O'Neill, P., and Moore, K. (2014). Quality control on the frontier. *Front. Genet.* 5:157. doi: 10.3389/fgene.2014.00157
- Pesquita, C., Ferreira, J. D., Couto, F. M., and Silva, M. J. (2014). The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources. *J. Biomed. Semant.* 5:4. doi: 10.1186/2041-1480-5-4
- Pisani, E., and AbouZahr, C. (2010). Sharing health data: good intentions are not enough. *Bull. World Health Organ.* 88, 462–466. doi: 10.2471/BLT.09.074393
- Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., et al. (2012). Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* 40, D940–D946. doi: 10.1093/nar/gkr972
- Sharma, M., Nunez-Garcia, J., Kearns, A. M., Doumith, M., Butaye, P. R., Argudín, M. A., et al. (2016). Livestock-associated methicillin resistant *Staphylococcus aureus* (LA-MRSA) clonal complex (CC) 398 isolated from UK animals belong to European lineages. *Front. Microbiol.* 7:1741. doi: 10.3389/fmicb.2016.01741
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255. doi: 10.1038/nbt1346
- Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., et al. (2005). Relations in biomedical ontologies. *Genome Biol.* 6:R46. doi: 10.1186/gb-2005-6-5-r46
- Tagini, F., Aubert, B., Troillet, N., Pillonel, T., Praz, G., Crisinel, P. A., et al. (2017). Importance of whole genome sequencing for the assessment of outbreaks in diagnostic laboratories: analysis of a case series of invasive *Streptococcus pyogenes* infections. *Eur. J. Clin. Microbiol. Infect. Dis.* doi: 10.1007/s10096-017-2905-z [Epub ahead of print].
- United Nations (2016). *Biodiversity and the 2030 Agenda for Sustainable Development*. Available at: <http://www.undp.org/content/undp/en/home/7010.html>
- librarypage/environment-energy/ecosystems_and_biodiversity/biodiversity-and-the-2030-agenda-for-sustainable-development---p.html
- van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., et al. (2014). A systematic review of barriers to data sharing in public health. *BMC Public Health* 14:1144. doi: 10.1186/1471-2458-14-1144
- Waldrum, A., Dolan, G., Ashton, P. M., Jenkins, C., and Dallman, T. J. (2017). Epidemiological analysis of *Salmonella* clusters identified by whole genome sequencing, England and Wales 2014. *Food Microbiol.* (in press). doi: 10.1016/j.fm.2017.02.012
- Wielinga, P. R., Hendriksen, R. S., Aarestrup, F. M., Lund, O., Smits, S. L., Koopmans, M. P., et al. (2017). “Global microbial identifier,” in *Applied Genomics of Foodborne Pathogens*, eds X. Deng, H. C. den Bakker, and R. S. Hendriksen (Cham: Springer International Publishing), 13–31.
- World Health Organization (2008). *Foodborne Disease Outbreaks : Guidelines for Investigation And Control*. Geneva: World Health Organization. Available at: <http://www.who.int/iris/handle/10665/43771>
- World Health Organization (2015). *WHO's First Ever Global Estimates of Foodborne Diseases Find Children Under 5 Account for Almost One Third of Deaths*. Geneva: World Health Organization. Available at: <http://www.who.int/mediacentre/news/releases/2015/foodborne-disease-estimates/en/>
- Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* 29, 415–420. doi: 10.1038/nbt.1823
- Zaidi, M. B., Calva, J. J., Estrada-Garcia, M. T., Leon, V., Vazquez, G., Figueroa, G., et al. (2008). Integrated food chain surveillance system for *Salmonella* spp. in Mexico. *Emerg. Infect. Dis.* 14, 429–435. doi: 10.3201/eid1403.071057

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Griffiths, Dooley, Graham, Van Domselaar, Brinkman and Hsiao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Food Safety in the Age of Next Generation Sequencing, Bioinformatics, and Open Data Access

Eduardo N. Taboada^{1,2}, Morag R. Graham^{1,3}, João A. Carriço⁴ and Gary Van Domselaar^{1,3*}

¹ National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada, ² Department of Biological Sciences, University of Lethbridge, Lethbridge, AB, Canada, ³ Department of Medical Microbiology and Infectious Diseases, Max Rady College of Medicine, University of Manitoba, Winnipeg, MB, Canada, ⁴ Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

OPEN ACCESS

Edited by:

Jennifer Ronholm,
McGill University, Canada

Reviewed by:

Andrey Tatarenkov,
University of California, Irvine,
United States

Badri Padukasahasram,
Illumina, United States
Mansel William Griffiths,
University of Guelph, Canada

*Correspondence:

Gary Van Domselaar
gary.vandomselaar@canada.ca

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 23 February 2017

Accepted: 04 May 2017

Published: 23 May 2017

Citation:

Taboada EN, Graham MR, Carriço JA and Van Domselaar G (2017) Food Safety in the Age of Next Generation Sequencing, Bioinformatics, and Open Data Access. *Front. Microbiol.* 8:909.
doi: 10.3389/fmicb.2017.00909

Public health labs and food regulatory agencies globally are embracing whole genome sequencing (WGS) as a revolutionary new method that is positioned to replace numerous existing diagnostic and microbial typing technologies with a single new target: the microbial draft genome. The ability to cheaply generate large amounts of microbial genome sequence data, combined with emerging policies of food regulatory and public health institutions making their microbial sequences increasingly available and public, has served to open up the field to the general scientific community. This open data access policy shift has resulted in a proliferation of data being deposited into sequence repositories and of novel bioinformatics software designed to analyze these vast datasets. There also has been a more recent drive for improved data sharing to achieve more effective global surveillance, public health and food safety. Such developments have heightened the need for enhanced analytical systems in order to process and interpret this new type of data in a timely fashion. In this review we outline the emergence of genomics, bioinformatics and open data in the context of food safety. We also survey major efforts to translate genomics and bioinformatics technologies out of the research lab and into routine use in modern food safety labs. We conclude by discussing the challenges and opportunities that remain, including those expected to play a major role in the future of food safety science.

Keywords: food safety, next-generation sequencing, genomic epidemiology, molecular typing, open data access

INTRODUCTION

The first complete sequence of a bacterial organism—*Haemophilus influenzae*—was generated in 1995, revealing for the first time the entire set of genetic information used to encode a free-living organism. Beyond the intrinsic scientific value of the 1.8 million base pair genome sequence and the nearly 1700 coding and non-coding genes within, this landmark scientific achievement is notable as the first demonstration that random shotgun sequencing combined with sophisticated computational methods can be used to successfully assemble a genome. The *H. influenzae* sequencing project also is notable for making the genome sequence data and the

bioinformatics software used to assemble it freely available to the scientific community. Such sharing aimed to be (and was) consistent with the policies initially set out by the ongoing Human Genome Project (NIH-DOE, 2012), later codified in the 1996 Bermuda Principles (Marshall, 2001). The policies of open sharing of genomic data and open source release of bioinformatics software tools set out by these seminal sequencing efforts were instrumental in cementing openness into the scientific culture (Lord et al., 2005). The impact of such open policies for scientific and medical advancement have been profound, as the publicly available genomic data and bioinformatics tools used to analyze these data are now routinely applied in nearly every aspect of biological and medical research, including the field of food safety science.

The fields of genomics and bioinformatics have been invaluable for advancing food safety science, although their application until recently has been limited toward research and development of molecular diagnostic technologies. For example, genomics and bioinformatics have been crucial in developing the standard molecular typing technologies currently in routine use for laboratory-based identification and tracking of foodborne disease outbreaks – namely Pulse-Field Gel Electrophoresis (PFGE), Multi-Locus Sequence Typing (MLST), and Multi-Locus Variable-Number of Tandem Repeats Analysis (MLVA). These tests require substantial bioinformatics and genomics to develop, but require only modest bioinformatics and genomics to carry out. What little bioinformatics and genomics that are required to conduct these tests historically have been incorporated within the various standardized lab procedures and software systems, effectively hidden “out of sight” (even “out of mind”) to most end users. This situation would radically change with the introduction of new, massively high throughput sequencing technologies, commonly referred to as Next-Generation Sequencing (NGS).

Next-generation sequencing was first made commercially available in late 2005 with the introduction of the GS20 sequencer manufactured by 454 Life Sciences. This new technology combined microfabrication advancements with an innovative new sequencing methodology to cheaply and rapidly generate massive amounts of nucleic acid sequencing data. Over the next decade, two main NGS technologies emerged, primarily distinguished by the sequence fragment (“read”) length generated. Short read technologies, such as those incorporated into the platform lines currently manufactured by Illumina and Life Technologies, generate read lengths from ~100 to ~600 bp with low per-base error rates (typically less than 1%) (Goodwin et al., 2016). These technologies are routinely used to assemble draft genome sequences containing multiple contiguous segments (contigs) with high accuracy and good coverage (>95% for an average bacterial genome). Two distinct longer read technologies are incorporated into the Pacific Biosystems (PacBio; Pacific Biosciences of California, Inc.) and Oxford Nanopore (Oxford Nanopore Technologies Ltd.) line of sequencers. The latter technologies both exploit single molecule sequencing to produce read lengths ranging from 1,000 to nearly 100,000 bp, although they still suffer from relatively high error rates (15–30%) (Goodwin et al., 2016). A current strength of long read technologies lies in their contribution

to generating “scaffolds” used for inter-connecting high quality contigs generated by short read technologies; in combination they permit efficient reconstruction of the draft genome. It is even possible (albeit expensive) to generate a high quality, “complete” bacterial draft genome using only long read technologies. Long read sequencing technologies have additional niche applications that may prove useful for future food science applications, as will be discussed below.

Perhaps the most important feature of NGS technologies is their ability to cheaply and quickly generate draft whole genome sequencing (WGS) data. This is especially true for microbial genomes, due to their smaller, more compact genomes relative to eukaryote genomes. The ability to routinely generate microbial draft genome sequence data has important applications in food safety science, particularly for foodborne disease surveillance and outbreak investigation. The conventional molecular typing technologies expose a mere fraction of the entire information contained within a foodborne pathogen’s genome, and thus provide limited ability to discriminate outbreak-related pathogen strains from unrelated, sporadically circulating strains. In contrast, WGS can theoretically reveal the entirety of the genome for a given microbial pathogen thereby allowing for the discrimination of strains that differ by a single nucleotide (amongst the millions of nucleotides comprising a typical bacterial pathogen genome). Although early pioneering studies applying WGS to outbreak analysis demonstrated much promise for this new technology, widespread recognition of its power would first occur in 2011.

GENOMICS AND BIOINFORMATICS IN THE LIMELIGHT

The ability of NGS technology to resolve the source of an outbreak was famously demonstrated during the 2010 Haiti cholera outbreak, the worst cholera epidemic in recent history killing at least 10,000 people and sickening well over 600,000 (Centers for Disease Control and Prevention [CDC], 2014). At the time of the outbreak, two hypotheses predominated as to its origin. One hypothesis argued that an endogenous pathogenic strain had been introduced from coastal waters; the other hypothesis suggested the cholera was introduced by UN peacekeepers deployed to Haiti after training in Kathmandu during a reported cholera outbreak spanning the country of Nepal (Maharjan, 2010). Conventional PFGE-based typing was insufficient to discriminate the outbreak strain from other environmental strains, and from other cholera outbreak strains originating mainly in Africa and Southeast Asia. The United States Centers for Disease Control and Prevention (CDC) performed NGS on a handful of strains from the Haitian outbreak and immediately released the data to the public. The free and open availability of this data allowed global researchers to compare the genome sequences from the Haitian strains with genome sequences from their own *Vibrio cholerae* collections, which they also rapidly released into the public domain (Chin et al., 2011; Hendriksen et al., 2011; Reimer et al., 2011). None of these early genomic epidemiological investigations were by

themselves sufficient to definitively trace the origin of the Haiti outbreak to the prior outbreak in Nepal; yet their combined genomic data, together with the available epidemiological data from the Haiti outbreak, provided overwhelming support to the “introduced outbreak strain” hypothesis that the outbreak was imported to Haiti from Nepal (Eppinger et al., 2014).

Shortly thereafter, a second major outbreak occurred that would have important consequences for food safety science: the 2011 Germany *Escherichia coli* O104:H4 outbreak. This large-scale outbreak of a novel strain of *E. coli* claimed over 50 lives and clinically affected a further 4,000 individuals (Grad et al., 2012). Following the example set by the Haiti cholera outbreak investigation teams, genome sequences for the O104 outbreak strains were immediately released to the public. The timing of the release coincided with the *Applied Bioinformatics for Public Health Microbiology* conference hosted at the Wellcome Trust Sanger Institute in Hinxton, United Kingdom, in the spring of 2011. The conference had assembled many of the world’s top bioinformatics scientists with expertise in microbial genomics, including the German researchers currently involved in the ongoing *E. coli* O104:H4 outbreak investigation. Using social media and other internet technologies, conference attendees joined with other researchers across the globe to perform the first crowdsourced, real-time analyses of the outbreak sequence data (Rasko et al., 2011). The *ad hoc* research group generated the outbreak pathogen’s draft genome sequence in under a day, and within a week they had designed molecular targets to distinguish the novel O104 outbreak strain from other circulating strains. Within that same short timeframe, they also determined the pathogen’s evolutionary origin and assessed its pathogenic potential (Boxrud et al., 2010; Chewapreecha et al., 2014). The extraordinary speed in which the novel O104 genome was characterized, largely as a result of the rapid public release of the pathogen genomic sequence data and its crowdsourced analysis, was widely reported in the scientific community (Mellmann et al., 2011; Owens, 2011; Rohde et al., 2011; Society for General Microbiology, 2011).

Beyond generating international headlines, these events, along with several other timely landmark genomic epidemiology investigations (Beres et al., 2010; Gilmour et al., 2010; Harris et al., 2010; Lewis et al., 2010; Gardy et al., 2011; Mutreja et al., 2011) spurred a grass-roots modernization movement. In the fall of 2012 an international consortium of scientists, clinicians, epidemiologists, and policy makers from public health, industry, medicine, and food regulatory sectors convened in Brussels to begin the process of planning out the global modernization of infectious disease diagnostics, surveillance, transmission, and outbreak investigation through adoption of NGS technologies (Aarestrup et al., 2012).

THE GLOBAL MICROBIAL IDENTIFIER CONSORTIUM

With accumulated evidence that NGS is more powerful than historical molecular subtyping methods, and fast becoming

more cost effective, pressures emerged to begin applying WGS for food safety. However, significant gaps remained to complicate widespread adoption of WGS: *For one, how would communication and multijurisdictional sharing of the large-scale WGS information be achieved for successful disease surveillance?* Fortunately the scientific community engaged early with public health, industry, clinicians, and food regulatory representatives to consider the broad needs of the global community. Such proactive, multi sector engagement and collaboration led to the creation of the Global Microbial Identifier (GMI) consortium (Wielinga et al., 2017), which envisions a global, interoperable analytical platform consisting of standardized pathogen genome databases, typing systems, and bioinformatics analysis tools for microbial and infectious disease identification, and diagnostics that will ultimately be made accessible to all nations with basic laboratory infrastructure (Global Microbial Identifier, 2017). Such an interoperable system should benefit not only the *One Health* frontlines at animal/human interfaces, but also food and agrifood industries, regulatory functions, policy makers, etc. Such a universally accessible platform also should benefit broader scientific, R&D and industrial applications.

The GMI vision is as challenging as it is ambitious. To clarify these challenges and develop a way forward, the GMI formed a number of working groups that have been instrumental in advancing genomics and bioinformatics for food safety: WG1 – *Political challenges, outreach and building a global network*; WG2 – *Repository and storage of sequence and associated metadata*; WG3 – *Analytical approaches*; and WG4 – *Methods validation, ring trials and proficiency assurance*. The manifold achievements and progress of these WGs are regularly updated at the GMI web site (Global Microbial Identifier, 2017); in this review we only focus on the activities that relate to open data, bioinformatics and food safety.

WG1: Political Challenges, Outreach, and Building a Global Network

From the beginning, the GMI WG1 recognized the extreme value of open access and integrated this philosophy as a core principle in its vision. It also appreciated that the adoption of such an approach will require global cooperation and coordination between many different and broad sectors, a large number of which have longstanding policies and laws governing data access and data sharing. In addition, many researchers working in these institutions hold provincial notions about the public health value of the data they possess, and thus are hesitant about rapid release of pathogen genomic data to the public archives. To achieve large-scale, global buy-in to the open data model, the concerns and needs of these stakeholders must be addressed. Thus, the focus of WG1’s activities includes identifying challenges and solutions regarding the varying sensitivities of metadata, intellectual property rights (IPR), and legal implications of open data as they apply to nations, regulatory agencies, and the food industry.

WG2: Repository and Storage of Sequence and Associated Metadata

GMI WG2 has been dealing with *Storage of sequence and associated contextual metadata*. The group advocates for rapid release of foodborne outbreak pathogen genomic data to the world's public archives. The group also promotes the standardization of the associated epidemiological, clinical, and laboratory metadata, for the purpose of facilitating data exchange and multijurisdictional approaches to outbreak control (Aarestrup and Koopmans, 2016).

To address concerns regarding the value of rapid release of standardized epidemiological metadata to the public domain versus the potential risk(s) that such information might expose to the institutions and nations contributing data, WG2 addressed the requisite issue of standardization. WG2 worked to enable a minimal common language for rapid release of pathogen genomic data that minimizes the legal risk of public data sharing while retaining the ability to conduct multinational outbreak investigations in real time (Aarestrup and Koopmans, 2016). Its solution exploits the fact that person-sensitive epidemiological data is not always required in order to detect emerging threats and outbreaks—contextual data (e.g., source country, year of isolation, origin, and whether (or not) it derived from an infection) are often sufficient. Consequently, such a minimum set of contextual data was developed (using controlled vocabularies) as the new MDM (or Minimal Data for Matching) reporting standard for data repository submissions of genome-scale pathogen sequence data (GMI meeting report 6). Both the US-hosted National Center for Biotechnology Information (NCBI)'s Short Read Archive (SRA) and the European Molecular Biology Laboratory (EMBL)'s European Nucleotide Archive (ENA) have adopted the GMI's MDM standard as minimal information fields to be reported for large-scale bacterial genome sequencing projects.

WG3: Analytical Approaches

GMI WG3 has been dealing with *Analytical approaches*, aiming to define the functional requirements of the major applications (e.g., typing, surveillance, diagnostics) into the global platform, and the analytical systems to be implemented to convert raw pathogen sequence data into actionable knowledge for public health and food regulatory response. WG3 has completed the mapping of the current analytical options and solutions against the needs of GMI end users; the group is currently developing systems for standardizing the comparison of different analytical pipelines. The group also has been active in developing benchmark datasets that can be used to validate the analytical pipelines as well as calibrating them to a common standard such that the results generated can be globally shared, compared, and consistently interpreted.

WG4: Methods Validation, Ring Trials, and Proficiency Assurance

GMI WG4 has endeavored to survey and promote partner lab consistency in both NGS data generation and data analyses, thereby ensuring that shared NGS data will remain high quality

and reliable. WG4 previously established a proficiency testing framework, and has run two full-sized, global proficiency tests focused on assessing the quality of partner lab sequencing of bacterial isolates and of control DNA, and of performing cluster analysis on sets of bacterial genome datasets (Moran-Gilad et al., 2015b; Reinert et al., 2015); PT2016 underway at time of writing). The early trials focused on the foodborne bacterial pathogens *E. coli* and *Salmonella enterica* Serovar Typhimurium; current trials are evaluating the foodborne pathogens *Listeria monocytogenes* and *Campylobacter* spp., and antimicrobial resistant *Klebsiella*. Future WG4 efforts aim to broaden analyses to include viral pathogens.

MODERNIZING FOOD SAFETY WITH GENOMICS, BIOINFORMATICS, AND OPEN DATA ACCESS

Several large-scale pilot projects have been implemented that apply NGS and modern bioinformatics analyses to existing foodborne disease surveillance programs. More specifically, these programs are aimed at replacing current subtyping approaches that underpin much of the modern food safety lab operations, with WGS data for real-time molecular surveillance. These modernization efforts represent one of the most crucial transformations in the history of food safety, with benefits and overall impact only starting to be realized. To get a sense of the scale required for this shift, it is important to review the role of molecular subtyping in infectious disease surveillance and control.

At its most basic level, subtyping is used to discriminate strains from the same species and to infer genetic relatedness, linking clinical cases representing a possible outbreak and further linking them to potential sources of infection (Sabat et al., 2013). More often, this goal is challenging to achieve amongst a background of sporadic cases in the absence of clear epidemiological links (Boxrud et al., 2010; Tauxe et al., 2010). Use of standardized laboratory protocols, standardized approaches for analysis and interpretation of data, and a common convention for naming molecular subtypes, collectively have been critical to large-scale deployment of subtyping for routine surveillance. The latter are best achieved in a public or open model, such as in the case of MLST where publicly available databases such as pubMLST (2017) are used by the global community as repositories for shared subtyping data, providing a means for efficient and open data exchange. A different model is sometimes necessary where, due to privacy concerns, restricted networks are required for secure data exchange. An example of this model has been the PulseNet network, which operates as an interconnected virtual laboratory network for the exchange of PFGE data by trusted members (namely laboratories of public health authorities and food safety regulators).

In considering WGS as a replacement for current subtyping methods, it is worth noting that in molecular epidemiology the key assumption is that subtyping data is a proxy for the underlying genomic information from which it is derived. Existing typing methods can thus be viewed as temporary

solutions in an era when rapid and inexpensive WGS was not possible; emergence of NGS and adoption of WGS are solving this limitation for public health and food safety investigations. Although WGS data can be analyzed using a traditional phylogenetic framework, the application of NGS in epidemiological surveillance requires approaches for WGS-based subtyping and additionally for relating WGS data to a subtype via a nomenclature scheme. WGS-subtyping facilitates efficient analysis of WGS data and is essential given the exponential increase in available data. A nomenclature is vital to the communication of results to public health or food safety professionals, allowing the monitoring of epidemiological trends and facilitating a rapid response aimed at disease prevention and control.

Of the two main strategies proposed for WGS-based subtyping, the first is based on the analysis of single nucleotide variant [SNV; also called single nucleotide polymorphism (SNP)] and small insertions/deletions (indels) between strains. Although this type of analysis can be performed on draft genome assemblies, several tools have been developed that directly compare raw sequence reads to a related reference genome sequence (Reinert et al., 2015). This process, which is referred to as variant detection by reference mapping, relies on algorithms that align each read to a reference genome and index the variation between them, also assigning confidence levels to each variant position based upon the sequence coverage and level of agreement between reads supporting the SNV (Mielczarek and Szyda, 2016). Reference mapping methodology has been used extensively in studies that have successfully used WGS in outbreak investigations (Harris et al., 2010; Gardy et al., 2011; Grad et al., 2012; Koser et al., 2012; Chewapreecha et al., 2014; Revez et al., 2014; Bekal et al., 2016). Reference mapping also is the approach that has been employed in analyzing *S. enterica* data within the large-scale, international GenomeTrakr project (Allard et al., 2016).

One of the challenges in reference mapping is that it is not always possible to identify an existing high-quality genome sufficiently similar to the genomes under study as a suitable reference genome. Although a closed and manually curated genome is preferable, it is feasible to apply a standard draft genome as the reference, provided that steps are taken to mask (filter out) regions posing problems for unambiguous read mapping (Lynch et al., 2016); these data can be generated on an *ad hoc* basis during the course of an investigation. Another challenge has been the development of nomenclature schemes for SNV reporting in the context of longitudinal pathogen surveillance. Recently, however, researchers at Public Health England have described an approach for systematically deriving pathogen subtype information based on a SNV-address approach (European Centre for Disease Prevention and Control, 2016). Moreover, because the SNV-based approach focuses on the subtlest form of genetic variation, it can be especially useful when investigating isolates exhibiting low levels of sequence variation, such as is expected when comparing outbreak-related isolates and investigating highly clonal or monomorphic populations (Machado et al., 2017).

The second major strategy for WGS-based subtyping is the ‘gene-by-gene’ approach, based on the original MLST concept (Maiden et al., 1998) but extended to the whole-genome level (wgMLST) (Sheppard et al., 2012; Maiden et al., 2013). MLST is based on indexing variation where each locus, a gene or gene fragment, is used as the basic unit of comparison. It has been proposed as a practical framework for developing hierarchical subtyping/nomenclature schemes suitable for studying strain relationships at a range of different resolution levels (Maiden et al., 2013). These include ribosomal MLST (rMLST) (Jolley et al., 2012), which targets 53 ribosomal protein subunit genes suitable for resolving bacterial isolates at all taxonomic levels; and core genome MLST (cgMLST) (Jolley and Maiden, 2010), which targets the genes shared by all or most members of a species (i.e., core genes). Genome-wide approaches to MLST have been applied to *Campylobacter jejuni* (Sheppard et al., 2012) and several other pathogens (Kohl et al., 2014; de Been et al., 2015; Moran-Gilad et al., 2015a; Pightling et al., 2015; Ruppitsch et al., 2015; Chen et al., 2016; Kluytmans-van den Bergh et al., 2016). The approach recently has been validated in a PulseNet International pilot project performing real-time NGS-based typing of *L. monocytogenes* (Jackson et al., 2016). PulseNet International also recently has committed to the wgMLST approach for their routine surveillance of foodborne disease (Carleton and Gerner-Smidt, 2016).

A drawback of cgMLST is that the numbers of genes in the core for any group of strains are dramatically lower than the total number available in a species ‘pan-genome,’ which is comprised of both the core and any accessory genes present in only some strains (Tettelin et al., 2008). It is possible, however, to design *ad hoc* MLST schemes based on the expanded number of genes shared by a smaller subset of genomes, thus providing additional discriminatory power when a low level of genetic variability is expected, as is the case in a rapidly expanding outbreak (Zhang et al., 2015). In addition, it is possible to extend the approach to whole genome MLST (wgMLST) by indexing allelic variation in both core and accessory genes. A hybrid analysis incorporating variation in core, accessory, and regulatory genome regions has recently been presented for the pathogenic *E. coli* lineage ST131 (McNally et al., 2016). Another potential problem with the gene-by-gene approach is that it collapses the diversity at multiple SNV sites located within a locus into a single allelic variant, greatly reducing discriminatory power. Nevertheless, species that are highly recombinogenic (essentially mosaic genomes) will benefit from this type of analytical treatment if the import of multiple SNVs in a single recombination event is a likely occurrence.

To permit stringent use of WGS data as standard public health practice, quality control metrics (such as sequence coverage) and interpretation criteria are needed. Regrettably, such metrics and criteria are still being defined for the field and remain a “moving target.” Additionally they do vary with bacterial species, the time frame of an investigation, and the methodology undertaken for the analyses; hence, no easy “one size fits all” approach exists. Although it remains premature to describe quality metrics and interpretation criteria in specific terms, key factors influencing sequence data generation have been

revealed (at least for the mature sequencing technologies) and there are ongoing global efforts to formalize how to generate reliable data and how to robustly interpret the data with confidence. The task is a difficult one since the sequencing parameters, timeframe of analysis, and evolutionary dynamics of the organism all influence the correlation of genomic variation and epidemiological interpretation in a complex way. Owing to the importance of data generation and interpretation in foodborne outbreak investigations, they remain high priorities and will receive considerable attention for the foreseeable future.

THE PROLIFERATION OF BIOINFORMATICS SOFTWARE FOR INFECTIOUS DISEASE ANALYSES

The shift in policy amongst food regulatory and public health agencies to rapidly release their pathogen genomic sequence data to the public allowed for the academic research community to join with government scientists, not just in the analysis of foodborne pathogen genomic data, but also in the creation of bioinformatics tools that can store, manage, and analyze the data. Additionally for the first time, genome sequences were being made publicly available for entire populations of pathogens, which spurred innovation in novel types of analyses performed, and in the development of “big data” approaches to efficiently analyze these vast datasets. The number and variety of bioinformatics software developed to analyze microbial data has grown tremendously, and is beyond any kind of comprehensive review. Here we report on some of the most popular and innovative bioinformatics software developed to tackle the analysis of large pathogen genome datasets with a focus on their application in food safety. For more in-depth reviews of the major bioinformatics pipelines used in foodborne disease surveillance, outbreak response, and diagnostics development, we refer the reader to the literature (Lynch et al., 2016; Ronholm et al., 2016).

Some of the first pipelines developed to facilitate analysis of large numbers of pathogen genomes were designed to generate phylogenies from whole genome sequence data. The variation identified among the analyzed sequences is used to infer phylogenetic trees providing supporting evidence for (or against) attributing a given isolate as part of an outbreak under study. In the previous section, we introduced two main approaches to capture this variation: SNV-based methods incorporated into pipelines such as the GenomeTrakr’s CFSAN SNP Pipeline (Davis et al., 2015); and gene-by-gene methods incorporated into whole genome MLST-based pipelines such as BIGSdb (Jolley and Maiden, 2010). A third approach, referred to as alignment-free methods, trades accuracy for speed in inferring the genetic distance between large populations of bacterial genomes and is useful for the rough clustering of thousands of genomes. One of the most notable implementations of this approach is Mash (Ondov et al., 2016), which can efficiently cluster upward of 50,000 draft bacterial genomes on a single CPU in just over a day.

A second major area of development is focused on *in silico* prediction of serotype for foodborne pathogens. These systems promise to drastically reduce cost and effort required to perform conventional antibody-based serotype determinations by instead predicting the serotype via analysis of the pathogen draft genome sequence. One such system, named SISTR (*Salmonella In Silico Typing Resource*) boasts an impressively high serovar predictive accuracy (~95%) (Yoshida et al., 2016). Additional serotype prediction systems have been built for other pathogens with demonstrated high predictive accuracy such as the SerotypeFinder system for *E. coli* (Joensen et al., 2015). It is expected that such systems will replace much of the conventional serotyping in food regulatory and public health labs.

A third field of active bioinformatics development is focused on the prediction of antimicrobial (antibiotic) resistance from NGS data. Some systems such as ResFinder (Kleinheinz et al., 2014) report the antimicrobial resistance genes they find in whole genome sequence data. ResFinder has high accuracy for finding antimicrobial resistance-associated genes, but cannot discriminate between allelic variants and their associated antibiotic resistant or sensitive phenotype. In contrast, the Comprehensive Antimicrobial Resistance Database (CARD) (Jia et al., 2017) incorporates curated models for each antimicrobial resistance gene and thus, can identify genes associated with antimicrobial resistance and also can predict whether they are resistant or sensitive to a given antibiotic or antibiotic class.

These powerful new bioinformatics tools hold great promise to augment or replace modern food safety lab tests and activities. However, to be used routinely in the front lines of foodborne disease surveillance and outbreak investigation, they need to be implemented in robust, user friendly software systems that shield the end user from the enormous complexity required to store, manage, and analyze vast amounts of data involved in these activities. Several commercial systems are available, such as Ridom SeqSphere+ (Ridom GmbH, Münster, Germany) and BioNumerics (Applied Maths, Sint-Martens-Latem, Belgium). These systems combine proprietary and open source analysis pipelines with sophisticated, easy-to-use interfaces that are familiar and intuitive for use by food safety investigators. One notable alternative is the completely open source Integrated Rapid Infectious Disease Analysis (IRIDA) platform, which provides a web-based end-to-end system for the storage, management, analysis, and sharing of NGS data (IRIDA, 2017). The IRIDA system is built to integrate multiple analytical pipelines in a common data storage and analysis system for genomic epidemiological applications. Other similar, albeit more focused, systems that provide easy-to-use interfaces with modern data analysis and visualization capacity include the Microreact system for phylogeographic analysis of SNV or MLST data (Argimón et al., 2016), the PHYLOViZ system (Ribeiro-Gonçalves et al., 2016; Nascimento et al., 2017) for epidemiological analysis and visualization of sequence (SNV and MLST) data, and GenGIS (Parks et al., 2009, 2013), which allows the overlay and analysis of phylogenetic data and associated metadata on digital maps.

WHAT IS NEXT? STANDARD VOCABULARIES FOR GENOMIC EPIDEMIOLOGY AND FOOD SAFETY

As mentioned above, the GMI:MDM standard (developed by the GMI and adopted by the world's public data archives) provides an important starting point for sharing the publicly available metadata. Yet much more can be achieved by having standards to describe the multiple layers of information associated with samples from microbial infection or food contamination events. This additional information, which is valuable for interpretation purposes, can range from the sample retrieval site (e.g., host-specific sites or environment) additional laboratory test results (e.g., antibiotic resistance profiles or additional typing methodologies), and possible clinical information (e.g., disease severity). The approaches developed to capture this information range from the definitions of a minimum information "checklist" to record the essential data, to fully-fledged ontologies, which provide a formal description of the entities in a given field of knowledge and the relationships between those entities. While their principal application is to create a machine-readable format that can be easily shared and understood between different databases and software, the process of constructing an ontology by domain experts allows the identification of the key concepts and steps that need to be described and shared. The best-known and most influential ontology in the field of molecular biology is the "Gene Ontology," which aims to provide a formal and descriptive representation of the biological function of genes (Ashburner et al., 2000). Its impact on biology has been profound: by providing a unifying tool that organizes and standardizes the staggering complexity of life, the Gene Ontology allows for the comprehensive analysis of biological function across all biological domains. Its success represents the impact that may be had by applying ontologies to other complex knowledge domains.

Since founding in 2005 and publication of its first seminal paper defining the Minimum Information about a Genome Sequence specification (MIGS) (Field et al., 2008), the Genomic Standards Consortium (GSC) (Genomic Standards Consortium, 2017) has been highly influential. The need to classify and annotate metagenome data also resulted in a refinement to MIGS to include metagenome metadata resulting in MIGS/MIMS (Garrity et al., 2008). Additional specifications such as Minimum Information about a Marker gene Sequence (MIMARKS) and Minimum Information about any (x) Sequence (MIXS) further refined their original standard (Yilmaz et al., 2011). Early GSC standards focused on sampling (geographic location, type of study) and sequencing information. More recently, the consortium has expanded the scope of their standardization efforts to include the environmental context of the biological identities sampled. This effort led to the development of the Environment Ontology (ENVO) that characterizes the sampling from general environmental sampling to specific body sites (Buttigieg et al., 2013).

However, extra effort to properly annotate sequence data and associated contextual metadata using standardized formats is still needed and additional field-specific information layers need to be directly applied to outbreak and population surveillance. Currently, the leading international effort in creating such a framework is being spearheaded by the Genomic Epidemiology Ontology (GenEpiO) consortium (GenEpiO, 2017). The GenEpiO consortium is tackling different aspects of the contextual metadata in order to facilitate the use of current or expanded ontology to genomic epidemiology investigations in clinical, food and environment surveillance and outbreaks. These range from defining specifications for reportable disease surveillance systems to standardizing food vocabularies, to more specific aspects of biological meaning such as describing antimicrobial resistance mechanisms. The latter requires updating and expansion of the Antimicrobial Resistance Ontology (ARO), developed by the Comprehensive Antibiotic Resistance Database (Jia et al., 2017), a manually curated repository of antimicrobial resistance mechanisms.

The need to annotate existing sequence-based microbial typing data also prompted the development of TypON, the microbial typing ontology (Vaz et al., 2014). This ontology focuses on the specification of sequence-based typing methods such as MLST, MLVA or single locus methodologies (e.g., *spa* typing for *S. aureus*, typing of the Short Variable Region (SVR) of Flagellin B for *Campylobacter* typing (Mellmann et al., 2004). The ontology is especially useful for annotating gene-by-gene methods (Sheppard et al., 2012) such as core or whole genome MLST, facilitating the comparison of existing schemas.

Although the application of NGS has brought great advances to epidemiological research over traditional methodologies for strain characterization, the whole process from the sample processing to sequencing and data analysis is more complex, leading to new challenges: multiple protocols for sample and library preparations are available and each sequencing run can use different versions of sequencing units and consumable reagents; moreover, in terms of data analysis, there are potentially hundreds of different software and respective versions that can be used and need to be tracked. Therefore, to facilitate comparative analyses the entire process from sample to analysis of results should be annotated. This need to capture process led to the creation of the Next Generation Sequencing Ontology (NGSOnto) (NGSOnto ontology, 2017). Using this ontological approach, researchers can maintain a description of the entire lab and data workflow, from sample collection to final results, thereby allowing for assessment of the experimental and bioinformatics pipelines for potential impacts on the resulting data interpretation.

All these efforts for standardization should contribute to a future in which data exchange may seamlessly occur, and truly interoperable resources may be created and shared (Sansone et al., 2012). Standardization and sharing will allow everyone globally to make the most use of the wealth of research and real world data that is being created via NGS technologies.

THE FUTURE IS BRIGHT FOR BIOINFORMATICS, GENOMICS, OPEN ACCESS, AND FOOD SAFETY

The first generation of genomic sequencing methods and analysis pipelines are now in the final stages of translation into routine application in modern food safety labs, and may soon replace many conventional laboratory tests. These advanced genomic and bioinformatics systems have proven their worth in reducing response times to emerging foodborne disease outbreaks, with substantial socioeconomic benefits in terms of improved public health, reduced health care costs, and avoidance of lost productivity due to illness (Scharff et al., 2016). The ongoing global efforts to modernize our food safety systems with genomics and bioinformatics have been impressive, but there remain many challenges and opportunities. Our current analytical capacity still requires the culturing of bacterial isolates, which can take several days. Culture-independent diagnostic testing using metagenomic technologies promises to do away with requisite culturing of isolates, thus shortening our response times even more. Culture independent metagenomics techniques have their own problems, however, such as the large amount of non-target data generated, contamination from environmental sources, and a current inability to distinguish between sequences derived from live or dead microorganisms. The vast and ever-growing size of the pathogen genome databases requires substantial high performance computing resources and novel algorithmic approaches to analyze such large data sets on a useful timescale. Addressing issues of data sharing and data ownership have only just begun. Metadata standardization is making good progress, but will require considerable sustained effort over multiple years to reach maturity. Many of the software pipelines and classification schemas still require extensive validation that will inevitably happen as more data is generated; hence we should anticipate some fluidity for both pipelines and schemas into the

future. Substantial effort will be required to increase our capacity to interpret this new type of comprehensive data and to train clinicians and epidemiologists in its use. As mentioned, quality control and quality analysis systems and metrics still need to be developed and standardized. Also, while open data access already has proven beneficial, it likely is not yet feasible for all food safety labs to achieve open data exchange in the immediate short term. However, mounting evidence is emerging for the public health and socioeconomic benefits of open access, plus the availability of bioinformatics tools and computing resources for all; and as these concepts are realized, they will drive a broader policy shift toward openness. Despite this daunting list of challenges that await resolution, implementation of genomics and bioinformatics technologies will occur, and without question will continue to transform our capacity to track and respond to foodborne disease threats.

AUTHOR CONTRIBUTIONS

All authors listed, have made equal, direct and intellectual contributions to the work, and approved it for publication.

FUNDING

This work was funded by the Public Health Agency of Canada (PHAC). The Agency had no role in design, opinions expressed or preparation of the manuscript.

ACKNOWLEDGMENT

We thank Dr. Celine Nadon for her helpful review of the manuscript.

REFERENCES

- Aarestrup, F. M., Brown, E. W., Detter, C., Gerner-Smidt, P., Gilmour, M. W., Harmsen, D., et al. (2012). Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. *Emerg. Infect. Dis.* 18:e1. doi: 10.3201/eid1811.120453
- Aarestrup, F. M., and Koopmans, M. G. (2016). Sharing data for global infectious disease surveillance and outbreak detection. *Trends Microbiol.* 24, 241–245. doi: 10.1016/j.tim.2016.01.009
- Allard, M. W., Strain, E., Melka, D., Bunning, K., Musser, S. M., Brown, E. W., et al. (2016). Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J. Clin. Microbiol.* 54, 1975–1983. doi: 10.1128/JCM.00081-16
- Argimón, S., Abudahab, K., Goater, R. J. E., Fedosejev, A., Bhai, J., Glasner, C., et al. (2016). Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genomics* 2:e000093. doi: 10.1099/mgen.0.000093
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556
- Bekal, S., Berry, C., Reimer, A. R., Van Domselaar, G., Beaudry, G., Fournier, E., et al. (2016). Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent *Salmonella enterica* serovar Heidelberg clone in the context of outbreak investigations. *J. Clin. Microbiol.* 54, 289–295. doi: 10.1128/JCM.02200-15
- Beres, S. B., Carroll, R. K., Shea, P. R., Sitkiewicz, I., Martinez-Gutierrez, J. C., Low, D. E., et al. (2010). Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4371–4376. doi: 10.1073/pnas.0911295107
- Boxrud, D., Monson, T., Stiles, T., and Besser, J. (2010). The role, challenges, and support of PulseNet laboratories in detecting foodborne disease outbreaks. *Public Health Rep.* 125(Suppl. 2), 57–62.
- Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., and Envo Consortium (2013). The environment ontology: contextualising biological and biomedical entities. *J. Biomed. Semantics* 4:43. doi: 10.1186/2041-1480-4-43
- Carleton, H. A., and Gerner-Smidt, P. (2016). Whole-genome sequencing is taking over foodborne disease surveillance. *Microbe* 11, 311–317.
- Centers for Disease Control and Prevention [CDC] (2014). *Cholera - Vibrio cholerae Infection. Cholera in Haiti.* Available at: <https://www.cdc.gov/cholera/haiti/> [updated November 07, 2014; accessed April 7, 2017].
- Chen, Y., Gonzalez-Escalona, N., Hammack, T. S., Allard, M. W., Strain, E. A., and Brown, E. W. (2016). Core genome multilocus sequence typing for identification of globally distributed clonal groups and differentiation of outbreak strains of *Listeria monocytogenes*. *Appl. Environ. Microbiol.* 82, 6258–6272. doi: 10.1128/AEM.01532-16

- Chewapreecha, C., Marttinen, P., Croucher, N. J., Salter, S. J., Harris, S. R., Mather, A. E., et al. (2014). Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet.* 10:e1004547. doi: 10.1371/journal.pgen.1004547
- Chin, C. S., Sorenson, J., Harris, J. B., Robins, W. P., Charles, R. C., Jean-Charles, R. R., et al. (2011). The origin of the haitian cholera outbreak strain. *N. Engl. J. Med.* 364, 33–42. doi: 10.1056/NEJMoa1012928
- Davis, S., Pettengill, J. B., Luo, Y., Payne, J., Shpunoff, A., Rand, H., et al. (2015). CFSAN SNP pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Comput. Sci.* 1:e20. doi: 10.7717/peerj.cs20
- de Been, M., Pinholt, M., Top, J., Bletz, S., Mellmann, A., van Schaik, W., et al. (2015). Core genome multilocus sequence typing scheme for high-resolution typing of *Enterococcus faecium*. *J. Clin. Microbiol.* 53, 3788–3797. doi: 10.1128/JCM.01946-15
- Eppinger, M., Pearson, T., Koenig, S. S., Pearson, O., Hicks, N., Agrawal, S., et al. (2014). Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized by the onset of the epidemic. *mBio* 5:e01721-14. doi: 10.1128/mBio.01721-14
- European Centre for Disease Prevention and Control (2016). *Multi-country Outbreak of Salmonella Enteritidis Phage Type 8 MLVA Type 2-9-7-3-2 Infections – First Update*. Stockholm: European Centre for Disease Prevention and Control.
- Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., et al. (2008). The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.* 26, 541–547. doi: 10.1038/nbt1360
- Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., et al. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* 364, 730–739. doi: 10.1056/NEJMoa1003176
- Garrison, G. M., Field, D., Kyripides, N., Hirschman, L., Sansone, S. A., Angiuoli, S., et al. (2008). Toward a standards-compliant genomic and metagenomic publication record. *OMICS* 12, 157–160. doi: 10.1089/omi.2008.A2B2
- GenEpiO (2017). *Genomic Epidemiology Ontology*. Available at: <http://genepio.org/> [updated March 9, 2017; accessed April 5, 2017].
- Genomic Standards Consortium (2017). *Genomic Standards Consortium*. Available at: <http://gensc.org/> [accessed February 25, 2017].
- Gilmour, M. W., Graham, M., Van Domselaar, G., Tyler, S., Kent, H., Trout-Yakel, K. M., et al. (2010). High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. *BMC Genomics* 11:120. doi: 10.1186/1471-2164-11-120
- Global Microbial Identifier (2017). *Global Microbial Identifier*. Available at: www.globalmicrobialidentifier.org [accessed February 25, 2017].
- Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49
- Grad, Y. H., Lipsitch, M., Feldgarden, M., Arachchi, H. M., Cerqueira, G. C., FitzGerald, M., et al. (2012). Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe 2011. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3065–3070. doi: 10.1073/pnas.1121491109
- Harris, S. R., Feil, E. J., Holden, M. T. G., Quail, M. A., Nickerson, E. K., Chantratita, N., et al. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327, 469–474. doi: 10.1126/science.1182395
- Hendriksen, R. S., Price, L. B., Schupp, J. M., Gillice, J. D., Kaas, R. S., Engelthaler, D. M., et al. (2011). Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *mBio* 2:e00157-11. doi: 10.1128/mbio.00157-11
- IRIDA (2017). *IRIDA – Integrated Rapid Infectious Disease Analysis Project*. Available at: <http://irida.ca> [updated February 28, 2017; accessed April 7, 2017].
- Jackson, B. R., Tarr, C., Strain, E., Jackson, K. A., Conrad, A., Carleton, H., et al. (2016). Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. *Clin. Infect. Dis.* 63, 380–386. doi: 10.1093/cid/ciw242
- Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., et al. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 45, D566–D573. doi: 10.1093/nar/gkw1004
- Joensen, K. G., Tetzschner, A. M., Iguchi, A., Aarestrup, F. M., and Scheutz, F. (2015). Rapid and easy in silico serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J. Clin. Microbiol.* 53, 2410–2426. doi: 10.1128/JCM.00008-15
- Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., et al. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiol. Read. Engl.* 158, 1005–1015. doi: 10.1099/mic.0.055459-0
- Jolley, K. A., and Maiden, M. C. (2010). BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 11:595. doi: 10.1186/1471-2105-11-595
- Kleinheinz, K. A., Joensen, K. G., and Larsen, M. V. (2014). Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage* 4:e27943. doi: 10.4161/bact.27943
- Kluytmans-van den Berg, M. F. Q., Rossen, J. W. A., Bruijning-Verhagen, P. C. J., Bonten, M. J. M., Friedrich, A. W., Vandebroucke-Grauls, C. M. J. E., et al. (2016). Whole-genome multilocus sequence typing of extended-spectrum-beta-lactamase-producing *Enterobacteriaceae*. *J. Clin. Microbiol.* 54, 2919–2927. doi: 10.1128/JCM.01648-16
- Kohl, T. A., Diel, R., Harmsen, D., Rothgänger, J., Walter, K. M., Merker, M., et al. (2014). Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. *J. Clin. Microbiol.* 52, 2479–2486. doi: 10.1128/JCM.00567-14
- Köser, C. U., Holden, M. T. G., Ellington, M. J., Cartwright, E. J. P., Brown, N. M., Ogilvy-Stuart, A. L., et al. (2012). Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N. Engl. J. Med.* 366, 2267–2275. doi: 10.1056/NEJMoa1109910
- Lewis, T., Loman, N. J., Bingle, L., Jumaa, P., Weinstock, G. M., Mortiboy, D., et al. (2010). High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. *J. Hosp. Infect.* 75, 37–41. doi: 10.1016/j.jhin.2010.01.012
- Lord, P., Macdonald, A., Sinnott, R., Ecklund, D., Westhead, M., and Jones, A. (2005). *Large-scale Data Sharing in the Life Sciences: Data Standards, Incentives, Barriers and Funding Models (“The Joint Data Standards Study”)*. Technical Report, No. UKEs-2006-02. Edinburgh: National e-Science Centre.
- Lynch, T., Petkau, A., Knox, N., Graham, M., and Van Domselaar, G. (2016). A primer on infectious disease bacterial genomics. *Clin. Microbiol. Rev.* 29, 881–913. doi: 10.1128/CMR.00001-16
- Machado, M. P., Ribeiro-Gonçalves, B., Silva, M., Ramirez, M., and Carriço, J. A. (2017). Epidemiological surveillance and typing methods to track antibiotic resistant strains using high throughput sequencing. *Methods Mol. Biol.* 1520, 331–355. doi: 10.1007/978-1-4939-6634-9_20
- Maharjan, L. (2010). *Cholera Outbreak Looms Over Capital*. *The Himalayan Times*. Kathmandu: International Media Network Nepal Pvt. Ltd.
- Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3140–3145. doi: 10.1073/pnas.95.6.3140
- Maiden, M. C. J., Jansen van Rensburg, M. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., et al. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* 11, 728–736. doi: 10.1038/nrmicro3093
- Marshall, E. (2001). Bermuda rules: community spirit, with teeth. *Science* 291, 1192. doi: 10.1126/science.291.5507.1192
- McNally, A., Oren, Y., Kelly, D., Pascoe, B., Dunn, S., Sreecharan, T., et al. (2016). Combined analysis of variation in core accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet.* 12:e1006280. doi: 10.1371/journal.pgen.1006280
- Mellmann, A., Harmsen, D., Cummings, C. A., Zentz, E. B., Leopold, S. R., Rico, A., et al. (2011). Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* 6:e22751. doi: 10.1371/journal.pone.0022751
- Mellmann, A., Mosters, J., Bartelt, E., Roggentin, P., Ammon, A., Friedrich, A. W., et al. (2004). Sequence-based typing of *flaB* is a more stable screening tool than

- typing of *flaA* for monitoring of *Campylobacter* populations. *J. Clin. Microbiol.* 42, 4840–4842. doi: 10.1128/JCM.42.10.4840-4842.2004
- Mielczarek, M., and Szyda, J. (2016). Review of alignment and SNP calling algorithms for next-generation sequencing data. *J. Appl. Genet.* 57, 71–79. doi: 10.1007/s13353-015-0292-7
- Moran-Gilad, J., Prior, K., Yakunin, E., Harrison, T. G., Underwood, A., Lazarovitch, T., et al. (2015a). Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. *Euro Surveill.* 20, 21186.
- Moran-Gilad, J., Sintchenko, V., Pedersen, S. K., Wolfgang, W. J., Pettengill, J., Strain, E., et al. (2015b). Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. *BMC Infect. Dis.* 15:174. doi: 10.1186/s12879-015-0902-3
- Mutreja, A., Kim, D. W., Thomson, N. R., Connor, T. R., Lee, J. H., Kariuki, S., et al. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477, 462–465. doi: 10.1038/nature10392
- Nascimento, M., Sousa, A., Ramirez, M., Francisco, A. P., Carriço, J. A., and Vaz, C. (2017). PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics* 33, 128–129. doi: 10.1093/bioinformatics/btw582
- NGSOnto ontology (2017). NCBO BioPortal - NGSOnto Ontology. Available at: <https://biportal.bioontology.org/ontologies/NGSONTO> [updated March 13, 2017; accessed April 5, 2017].
- NIH-DOE (2012). National Human Genome Research Institute (NHGRI) - Access to Mapping and Sequencing Resources. NIH-DOE Guidelines for Access to Mapping and Sequencing Data and Material Resources. Available at: <http://www.genome.gov/10000925> [updated March 9, 2012; accessed April 5, 2017].
- Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x
- Owens, B. (2011). The German *E. coli* Outbreak: 40 Lives and Hours of Crowdsourced Sequence Analysis Later. *Nature News Blog*. Available at: http://blogs.nature.com/news/2011/06/the_german_e_coli_outbreak_40.html [updated June 20, 2011; accessed February 27, 2017].
- Parks, D. H., Mankowski, T., Zangoei, S., Porter, M. S., Armanini, D. G., Baird, D. J., et al. (2013). GenGIS 2: geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework. *PLoS ONE* 8:e69885. doi: 10.1371/journal.pone.0069885
- Parks, D. H., Porter, M., Churcher, S., Wang, S., Blouin, C., Whalley, J., et al. (2009). GenGIS: a geospatial information system for genomic data. *Genome Res.* 19, 1896–1904. doi: 10.1101/gr.095612.109
- Pightling, A. W., Petronella, N., and Pagotto, F. (2015). The *Listeria monocytogenes* core-genome sequence typer (LmCGST): a bioinformatic pipeline for molecular characterization with next-generation sequence data. *BMC Microbiol.* 15:224. doi: 10.1186/s12866-015-0526-1
- pubMLST (2017). pubMLST. Available at: <http://www.pubmlst.org> (accessed February 27, 2017).
- Rasko, D. A., Webster, D. R., Sahl, J. W., Bashir, A., Boisen, N., Scheutz, F., et al. (2011). Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.* 365, 709–717. doi: 10.1056/NEJMoa1106920
- Reimer, A. R., Van Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., et al. (2011). Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. *Emerg. Infect. Dis.* 17, 2113–2121. doi: 10.3201/eid1711.110794
- Reinert, K., Langmead, B., Weese, D., and Evers, D. J. (2015). Alignment of next-generation sequencing reads. *Annu. Rev. Genomics Hum. Genet.* 16, 133–151. doi: 10.1146/annurev-genom-090413-025358
- Revez, J., Llarena, A.-K., Schott, T., Kuusi, M., Hakkinen, M., Kivistö, R., et al. (2014). Genome analysis of *Campylobacter jejuni* strains isolated from a waterborne outbreak. *BMC Genomics* 15:768. doi: 10.1186/1471-2164-15-768
- Ribeiro-Gonçalves, B., Francisco, A. P., Vaz, C., Ramirez, M., and Carriço, J. A. (2016). PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. *Nucleic Acids Res.* 44, W246–W251. doi: 10.1093/nar/gkw359
- Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N. J., Hentschke, M., et al. (2011). Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N. Engl. J. Med.* 365, 718–724. doi: 10.1056/NEJMoa1107643
- Ronholm, J., Nasheri, N., Petronella, N., and Pagotto, F. (2016). Navigating microbiological food safety in the era of whole-genome sequencing. *Clin. Microbiol. Rev.* 29, 837–857. doi: 10.1128/CMR.00056-16
- Ruppitsch, W., Pietzka, A., Prior, K., Bletz, S., Fernandez, H. L., Allerberger, F., et al. (2015). Defining and evaluating a core genome MLST scheme for whole genome sequence-based typing of *Listeria monocytogenes*. *J. Clin. Microbiol.* 53, 2869–2876. doi: 10.1128/JCM.01193-15
- Sabat, A. J., Budimir, A., Nashev, D., Sá-Leão, R., van Dijl, J. M., Laurent, F., et al. (2013). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveill.* 18:20380.
- Sansone, S. A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., et al. (2012). Toward interoperable bioscience data. *Nat. Genet.* 44, 121–126. doi: 10.1038/ng.1054
- Scharff, R. L., Besser, J., Sharp, D. J., Jones, T. F., Peter, G. S., and Hedberg, C. W. (2016). An economic evaluation of PulseNet: a network for foodborne disease surveillance. *Am. J. Prev. Med.* 50, S66–S73. doi: 10.1016/j.amepre.2015.09.018
- Sheppard, S. K., Jolley, K. A., and Maiden, M. C. J. (2012). A gene-by-gene approach to bacterial population genomics: whole genome MLST of *Campylobacter*. *Genes* 3, 261–277. doi: 10.3390/genes3020261
- Society for General Microbiology (2011). Crowd-sourcing the *E. coli* O104:H4 Outbreak. *Science Daily*. Available at: <http://www.sciencedaily.com/releases/2011/09/110904215952.htm> [updated September 6, 2011; accessed February 27, 2017].
- Tauxe, R. V., Doyle, M. P., Kuchenmüller, T., Schlundt, J., and Stein, C. E. (2010). Evolving public health approaches to the global challenge of foodborne infections. *Int. J. Food Microbiol.* 139(Suppl. 1), S16–S28. doi: 10.1016/j.ijfoodmicro.2009.10.014
- Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* 11, 472–477. doi: 10.1016/j.mib.2008.09.006
- Vaz, C., Francisco, A. P., Silva, M., Jolley, K. A., Bray, J. E., Pouselee, H., et al. (2014). TypOn: the microbial typing ontology. *J. Biomed. Semantics* 5:43. doi: 10.1186/2041-1480-5-43
- Wielinga, P. R., Hendriksen, R. S., Aarestrup, F. M., Lund, O., Smits, S. L., Koopmans, M. P. G., et al. (2017). “Global microbial identifier,” in *Applied Genomics of Foodborne Pathogens*, eds X. Deng, H. C. den Bakker, and R. S. Hendriksen (Cham: Springer International Publishing), 13–31. doi: 10.1007/978-3-319-43751-4_2
- Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* 29, 415–420. doi: 10.1038/nbt.1823
- Yoshida, C. E., Kruczakiewicz, P., Laing, C. R., Lingohr, E. J., Gannon, V. P., Nash, J. H., et al. (2016). The *Salmonella* in silico typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS ONE* 11:e0147101. doi: 10.1371/journal.pone.0147101
- Zhang, J., Halkilahti, J., Hänninen, M.-L., and Rossi, M. (2015). Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy. *J. Clin. Microbiol.* 53, 1765–1767. doi: 10.1128/JCM.00051-15
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Copyright © 2017 Taboada, Graham, Carriço and Van Domselaar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The Public Health Impact of a Publically Available, Environmental Database of Microbial Genomes

Eric L. Stevens ^{*}, Ruth Timme, Eric W. Brown, Marc W. Allard, Errol Strain, Kelly Bunning and Steven Musser

Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, MD, USA

Keywords: whole-genome sequencing, databases, genetic, foodborne pathogens, foodborne illness prevention, public health

OPEN ACCESS

Edited by:

Sabah Bidawid,
Health Canada, Canada

Reviewed by:

Giovanna Suzzi,
University of Teramo, Italy
Lorenza Putignani,
Bambino Gesù Ospedale Pediatrico
(IRCCS), Italy

***Correspondence:**

Eric L. Stevens
eric.stevens@fda.hhs.gov

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 07 December 2016

Accepted: 19 April 2017

Published: 09 May 2017

Citation:

Stevens EL, Timme R, Brown EW, Allard MW, Strain E, Bunning K and Musser S (2017) The Public Health Impact of a Publically Available, Environmental Database of Microbial Genomes. *Front. Microbiol.* 8:808.
doi: 10.3389/fmicb.2017.00808

Imagine a public health resource that contains the whole-genomic sequences of tens of thousands of microbial pathogens that can be accessed by anyone in the world at any time and without cost or registration. Furthermore, each of these stored genetic sequences also has a wealth of metadata attached to it that provides additional and useful identifying information. From providing the geographic location of where the food or environmental isolate was collected to identifying the year that the isolate was obtained, this combination of genomic information and its accompanying and descriptive information (i.e., metadata) can be used to inform the interpretation of phylogenetic trees constructed using the sequence data. Finally, now imagine that this process is automated and each day the phylogenetic tree of a specific species is updated, allowing public health scientists to infer the evolutionary history and relationship of relevant isolates to not only resolve foodborne outbreaks, but to prevent them all together. One may next think to ask how far are we away from realizing this vision, and the answer is that we are already there.

The underlying science behind using whole-genome sequencing (WGS) data to link clinical isolates back to its environmental or contaminated food source is simple: is the DNA of the environmental or food isolate and the clinical isolate genetically related to each other? That is, are the 3–5 million nucleotides that make up the whole-genomic sequences of those isolates under consideration identical or do they differ by only a few nucleotide changes [e.g., A to G mutation; called a single nucleotide polymorphism (SNP)]. Since foodborne pathogens often have short generation times under optimal growing conditions, it is expected that a small number of mutations, or SNPs, will be acquired from the time of product contamination to the isolation of the clinical specimen. Therefore, there could be a range as to the number of SNP differences seen among clinical, food, and environmental isolates that are all part of an outbreak, but these can still easily cluster together and even be distinct from non-outbreak related isolates. That is the power and utility of using genomic sequence data.

However, using genetic information for identification purposes is not a novel concept. In fact, much of how bacterial DNA is currently being used across various food safety applications is based on similar methods that have been employed using human or viral genetic information. For instance, comparing the DNA sample collected at a crime scene and comparing that to the DNA of multiple suspects is very similar to comparing a bacterial isolate derived from a clinical patient and comparing that to different environmental or food isolates that could have led to the cause of the clinical illness. Matching of human DNA has also been used to determine the paternity of a child for where there is more than one potential father, which is also very similar to determining the source of an environmental isolate from a production facility to one of several

possible ingredients that could have introduced the contamination into the facility from an earlier part in the supply chain. This approach has also been applied to delimit both the transmission and spread of viral outbreaks, including HIV (Ou et al., 1992), Ebola (Holmes et al., 2016), and Cholera (Chin et al., 2011), much in the same way that it is currently being used as a real-time molecular epidemiological tool for foodborne disease surveillance. Furthermore, WGS has also been used to track antimicrobial-resistant bacterial infections in real-time in a hospital setting (Snitkin et al., 2012).

Perhaps one of the most promising applications of building a microbial reference database filled with the genomic sequences of environmental and food isolates collected from regions all around the world is almost identical to how human DNA sequences can reveal from which countries a person's ancestors most likely came from. To think that combining microbial DNA information and its geographic source can speed up outbreak investigations by suggesting potential sources is just one of the many public health benefits of utilizing this advancing technology. This application and use is more than just promising; it is essential considering the increasingly global food supply and its resulting supply chain that feeds both domestic and international populations. Indeed, the application of this technology for foodborne outbreak investigations has already begun to reduce the time required to identify the source of the outbreak. By narrowing the epidemiological focus using WGS' increased resolution over traditional subtyping methods (e.g., PFGE) in differentiating between outbreak and non-outbreak related samples, less time is spent tracking the outbreak to its source in order to remove it from the food supply.

However, this technology also has the potential to severely limit or prevent outbreaks from occurring when used as part of preventative controls, which is the focus of the recent Food Safety and Modernization Act (FSMA; Pub. L. 111-353). Routine environmental and product monitoring by both industry and regulatory agencies has already been used to identify and link a contaminated product to a clinical illness very early on during the course of an outbreak, allowing for a quick response that led to a recall of smaller amounts of the contaminated product. While no one can be certain, it is likely that these actions prevented many other individuals from becoming sick and likely saved the company more than if a greater amount of the contaminated product had made its way to market. Decreasing the amount of recalled product, as well as reducing the number of sick individuals saves money both in terms of fewer lawsuits and protecting brand recognition.

Further, uses of this technology within production facilities have demonstrated its ability to resolve between transient and resident pathogens, giving industry a powerful, and precise tool to track and trace sources of contamination by themselves as part of their requirements for environmental monitoring under FSMA. Companies are already routinely employing WGS to resolve contaminating niche locations within their production lines, preventing their finished products from ever becoming contaminated. It is this use of WGS that will perhaps have the

biggest impact on public health in the future by significantly reducing the number of contaminated products entering the market, thereby decreasing both the frequency and size of foodborne outbreaks by making the food supply safer.

Regardless of whether WGS is used within a preventative control or outbreak response framework, its success is predicated on two things: (1) using the genomic sequences of isolates to differentiate between related and unrelated samples; and (2) having a sufficient number of reference isolates from which to compare against. For preventative controls, only the isolates specific to that supply chain (e.g., environmental sampling or isolates from ingredients coming in or out of the production facility) are necessary, and these databases could be private or industry-specific. However, if outbreak detection or tracking isolates across the global supply chain is the focus of WGS technology, then it is paramount to have a freely-available database filled with environmental and food isolates from all over the world that anyone can access.

And that is precisely what the United States Food and Drug Administration (FDA) realized at the same time that the cost of WGS began to rapidly decrease in the late 2000s. Therefore, in 2011 FDA launched GenomeTrakr (FDA, 2016), a genomic repository for storing the sequences of food or environmental isolates and housed at the National Center for Biotechnology Information (NCBI, 2016). This distributed network of federal, state, international, and public health laboratories houses mainly food and environmental isolates and combines this genomic information with an isolate's descriptive metadata. As previously described, this combination serves as an important resource for both public health agencies and industry to compare the genetic sequences of both clinical and environmental isolates against.

GenomeTrakr currently contains this information for more than 70,000 isolates, and continues to grow at a remarkable pace as more states and countries understand the benefit of WGS technology for food safety (see **Figure 1**). With new laboratories and partners joining the network each month, GenomeTrakr's expanding reference library is more likely to house a sequenced isolate that is genetically related to a genome that is queried against those 70,000 isolates. This ability to provide for a possible genetic match only becomes more probable as environmental and product sampling increases across the numerous production regions around the world. In other words, the geographic information attached to each isolate is then better able to supply critical information helpful in resolving the myriad possible sources of contamination if many isolates are being sampled from a uniform distribution (i.e., isolates come from many states, countries, and from a wide array of food commodities that represent variation seen among the different bacterial species).

It is inevitable that the number of sequenced environmental and food isolates will only continue to increase as more and more countries and industries embrace WGS technology its position as a single microbiological test (e.g., serotyping, antimicrobial resistance, pathogenicity, etc.). These genomic sequences—and the associated metadata—need to be made available for its successful use as a public health resource. However, there are

Bacterial Sequences in GenomeTrakr Database

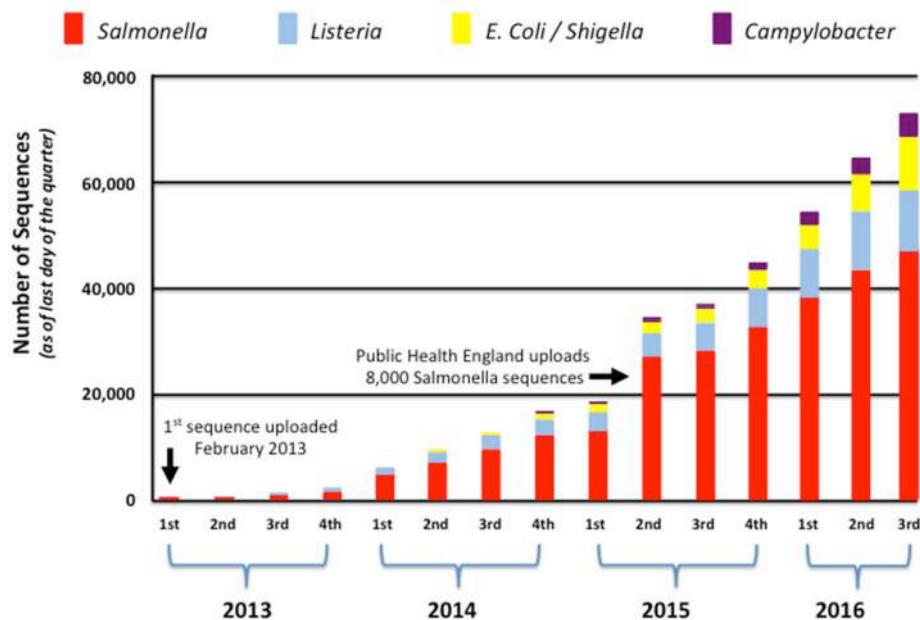


FIGURE 1 | Growth of bacterial genomes within GenomeTrakr (2013–2016). This figure shows the number of bacterial genomes that have been added to FDA's GenomeTrakr at the end of each quarter spanning 2013–2016. Four pathogens (*Salmonella*, *Listeria*, *E. coli / Shigella*, and *Campylobacter*) are the dominant foodborne pathogens collected and made available to the public.

serious concerns that arise from the sharing of such potentially sensitive data, especially when it could be used to self-implicate industry if supplied environmental or product samples match to clinical isolates. On the other hand, routine use of this technology by industry within their own production lines and supply chains can help mitigate that problem by becoming aware of and then correcting the contamination internally before it is able to result in a clinical illness. Industry could also use this technology to identify possible sources of contamination within their supply chain, especially if they receive raw ingredients from multiple and distinct sources. There are also concerns that this technology could adversely affect developing or transitional countries, and even smaller companies that lack the necessary resources able to robustly utilize this technology for their own purposes of surveillance and preventative controls. It is hoped that all stakeholders will continue to engage in meaningful dialogue to ensure the most beneficial use of this technology for public health while balancing private and commercial interests.

Nevertheless, environmental sampling and the availability of an isolate's genomic information is fundamental in order to gain an understanding into the inherent geographic variation and spread of potential foodborne pathogens coming out of

the fields and at various points along the supply chain. The public health benefit of providing public access to this data will only become more vital as our ability to process and analyze this data become more sophisticated. Indeed, GenomeTrakr has already successfully demonstrated how a large database of genomic sequences and metadata can be used effectively for food safety within the United States by not only helping to decrease the length of foodborne outbreak investigations but by detecting outbreaks earlier. Going forward, GenomeTrakr can serve as a model for a large database of environmental isolates from all over the world that can now be utilized for global public health benefit, and it is imagined that this resource will be beneficial to both public health agencies and industry.

AUTHOR CONTRIBUTIONS

Wrote the paper: ES, RT, MA, ES, KB, SM, and EB.

FUNDING

This work was supported by the Center for Food Safety and Applied Nutrition at the U.S. Food and Drug Administration.

REFERENCES

- Chin, C. S., Sorenson, J., Harris, J. B., Robins, W. P., Charles, R. C., Jean-Charles, R. R., et al. (2011). The origin of the Haitian cholera outbreak strain. *N Engl J Med.* 364, 33–42. doi: 10.1056/NEJMoa1012928
- FDA (2016). *GenomeTrakr*. Available online at: <http://www.fda.gov/Food-FoodScienceResearch/WholeGenomeSequencingProgramWGS/ucm363134.htm>
- Holmes, E. C., Dudas, G., Rambaut, A., and Andersen, K. G. (2016). The evolution of Ebola virus: insights from the 2013–2016 epidemic. *Nature* 538, 193–200. doi: 10.1038/nature19790
- NCBI (2016). *Pathogen Detection*. Available online at: <http://www.ncbi.nlm.nih.gov/pathogens/>
- Ou, C. Y., Ciesielski, C. A., Myers, G., Bandea, C. I., Luo, C. C., Korber, B. T., et al. (1992). Molecular epidemiology of HIV transmission in a dental practice. *Science* 256, 1165–1171. doi: 10.1126/science.256.5060.1165
- Snitkin, E. S., Zelazny, A. M., Thomas, P. J., Stock, F., Group, N. C. S. P., Henderson, D. K., et al. (2012). Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci. Transl. Med.* 4:148ra116. doi: 10.1126/scitranslmed.3004129

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Stevens, Timme, Brown, Allard, Strain, Bunning and Musser. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens

Lee S. Katz^{1,2*}, Taylor Griswold^{1,3}, Amanda J. Williams-Newkirk^{1,4}, Darlene Wagner^{1,4}, Aaron Petkau⁵, Cameron Sieffert⁵, Gary Van Domselaar⁵, Xiangyu Deng² and Heather A. Carleton¹

¹ Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, GA, USA, ² Center for Food Safety, College of Agricultural and Environmental Sciences, University of Georgia, Griffin, GA, USA, ³ Oak Ridge Institute for Science and Education, Oak Ridge Associated Universities, Oak Ridge, TN, USA, ⁴ IHRC, Inc., Atlanta, GA, USA, ⁵ National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada

OPEN ACCESS

Edited by:

Sandra Torriani,
University of Verona, Italy

Reviewed by:

Jason Sahl,
Northern Arizona University, USA
Young Min Kwon,
University of Arkansas, USA

*Correspondence:

Lee S. Katz
gzu2@cdc.gov

Specialty section:

This article was submitted to
Food Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 16 December 2016

Accepted: 23 February 2017

Published: 13 March 2017

Citation:

Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, Van Domselaar G, Deng X and Carleton HA (2017) A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. *Front. Microbiol.* 8:375.
doi: 10.3389/fmicb.2017.00375

Modern epidemiology of foodborne bacterial pathogens in industrialized countries relies increasingly on whole genome sequencing (WGS) techniques. As opposed to profiling techniques such as pulsed-field gel electrophoresis, WGS requires a variety of computational methods. Since 2013, United States agencies responsible for food safety including the CDC, FDA, and USDA, have been performing whole-genome sequencing (WGS) on all *Listeria monocytogenes* found in clinical, food, and environmental samples. Each year, more genomes of other foodborne pathogens such as *Escherichia coli*, *Campylobacter jejuni*, and *Salmonella enterica* are being sequenced. Comparing thousands of genomes across an entire species requires a fast method with coarse resolution; however, capturing the fine details of highly related isolates requires a computationally heavy and sophisticated algorithm. Most *L. monocytogenes* investigations employing WGS depend on being able to identify an outbreak clade whose inter-genomic distances are less than an empirically determined threshold. When the difference between a few single nucleotide polymorphisms (SNPs) can help distinguish between genomes that are likely outbreak-associated and those that are less likely to be associated, we require a fine-resolution method. To achieve this level of resolution, we have developed Lyve-SET, a high-quality SNP pipeline. We evaluated Lyve-SET by retrospectively investigating 12 outbreak data sets along with four other SNP pipelines that have been used in outbreak investigation or similar scenarios. To compare these pipelines, several distance and phylogeny-based comparison methods were applied, which collectively showed that multiple pipelines were able to identify most outbreak clusters and strains. Currently in the US PulseNet system, whole genome multi-locus sequence typing (wgMLST) is the preferred primary method for foodborne WGS cluster detection and outbreak investigation due to its ability to name standardized genomic profiles, its central database, and its ability to be run in a graphical user

interface. However, creating a functional wgMLST scheme requires extended up-front development and subject-matter expertise. When a scheme does not exist or when the highest resolution is needed, SNP analysis is used. Using three *Listeria* outbreak data sets, we demonstrated the concordance between Lyve-SET SNP typing and wgMLST.

Availability: Lyve-SET can be found at <https://github.com/lskatz/Lyve-SET>.

Keywords: SNP pipeline, wgMLST, genomic epidemiology, foodborne, outbreak, bacterial pathogen

INTRODUCTION

Modern outbreak investigation is enhanced with molecular subtyping evidence. These lines of evidence have been, but are not limited to: pulsed-field gel electrophoresis (PFGE), multiple-locus variable number tandem repeat analysis (MLVA), and multi-locus sequence typing (MLST; MacCannell, 2013). Each of these methods yields specific targets to measure genetic relatedness among pathogens isolated from human cases, animals, foods, or the environment, resulting in evidence for or against their inclusion in a cluster, which in turn aids in epidemiological investigations. In the age of whole genome sequencing (WGS), outbreak investigation is being increasingly supported by phylogenomic methods that are more robust and discriminatory than any aforementioned subtyping method (Jackson et al., 2016). Whether infectious disease outbreaks are caused by single pathogenic clones or by multiple clones, a basic assumption can be made that the epidemiological association between cases can be inferred from the phylogenetic relationships between the case-defining microorganisms. In an outbreak scenario as phylogenetic relatedness increases, the likelihood of epidemiological concordance increases. In other words, phylogeny approximates epidemiology.

There are two dominant methods to create phylogenies for WGS-enhanced outbreak investigations: whole-genome multi-locus sequence typing (wgMLST) and single nucleotide polymorphisms (SNPs). In the wgMLST method for a single genome, as in conventional MLST (Maiden et al., 1998), loci are compared against a database of known alleles and either labeled with a known allele identifier or given a new allele identifier. In MLST and wgMLST, alleles are either the same or different, meaning that any single nucleotide substitution, insertion, or deletion equates to an allele change. With wgMLST, thousands of loci are compared and their distances are used to generate a phylogeny usually with either the unweighted-pair-group-method-with-arithmetic-mean (UPGMA) or neighbor-joining (NJ) algorithm. One implementation of wgMLST is through the BioNumerics software (Applied Maths, Sint-Martens-Latem, Belgium).

In the SNP-based method, single nucleotide changes are used to infer phylogenetic relatedness. This method is implemented in many software packages. Snp-Pipeline has been used for regulatory evidence of *Salmonella enterica* at the Center for Food Safety and Applied Nutrition (CFSAN; Pettengill et al., 2014; Davis et al., 2015). RealPhy has been used to characterize *Clostridium botulinum* outbreaks (Bertels et al., 2014; Shirey et al., 2016). SNVPhyl is used by the National Microbiology Laboratory (NML) of the Public Health Agency of Canada

(PHAC) for, among other organisms, *S. enterica* (Bekal et al., 2016). Most SNP-based methods have a common workflow: (1) mapping raw reads onto a reference genome, (2) identifying SNPs, (3) removing lower-quality SNPs, (4) creating a multiple sequence alignment (MSA) from selected SNPs, and (5) inferring a phylogeny from the MSA. When these SNP-based methods remove SNPs with less support, they can be called high-quality SNP-based methods (hqSNP). SNPs with less support can be identified by having few raw reads, by having conflicting allele calls in the raw reads (i.e., low consensus), by occurring in mutation hotspot regions such as phage regions, or for many other reasons. A modification of this typical SNP workflow is implemented in kSNP where nucleotides of odd length k (k -mers) are extracted from raw reads supplied for the genomes being analyzed (Gardner et al., 2015). Instead of aligning to a reference, the k -mers from one genome are compared against the other genome's, where the middle nucleotide can be variable. These variable bases can be extracted into a pseudo-multiple sequence alignment such that a phylogeny can be built. kSNP has been used in describing the population structure of certain foodborne pathogens, e.g., *L. monocytogenes* in cured ham in Italy (Morganti et al., 2015), and in outbreak investigations of other bacterial pathogens, e.g., retrospective analysis of *Legionella pneumophila* (Mercante et al., 2016). These hqSNP pipelines increase the signal-to-noise ratio in favor of a high-quality phylogeny at the risk of removing true but low-quality SNPs.

In 2013, the NML and the Enteric Diseases Laboratory Branch (EDLB) of The Centers for Disease Control and Prevention (CDC) briefly described an initial SNP-based workflow called the SNP Extraction Tool (SET). The initial version of SET was used for the Haiti cholera outbreak of 2010 (Katz et al., 2013). The common code base of SET has since been forked, with the NML branch rebranded as SNVPhyl (Petkau et al., 2016) and the CDC version as Lyve-SET, named after the organisms with which it was first used: *Listeria*, *Yersinia*, *Vibrio*, and Enterobacteriaceae. Since 2013, the Centers of Disease Control and Prevention (CDC) has participated in an interagency collaboration to routinely sequence and analyze all clinical and food-related *Listeria monocytogenes* isolates in the US with the eventual goal to replace PFGE (Carleton and Gerner-Smith, 2016). As WGS data of these isolates are being continuously generated, a phylogenetic framework needs to be constructed and constantly updated to support epidemiological surveillance and outbreak investigation of *L. monocytogenes*. Therefore, upon the onset of the interagency collaboration, we revised and formalized Lyve-SET into a packaged pipeline that suits the needs of bacterial foodborne outbreak investigations. Lyve-SET was refined in the context of *L. monocytogenes* outbreak investigations and

continues to be a strong reference tool for *L. monocytogenes* and many other foodborne pathogens such as *S. enterica*, *Escherichia coli*, *Yersinia enterocolitica*, *Cronobacter*, and *Vibrio cholerae*.

Historically, it has been difficult to evaluate and compare SNP pipelines, and an even bigger challenge to compare them to workflows based on other algorithms (e.g., wgMLST). Each of the aforementioned pipelines produces output that can be used for interpreting the relationship between genomes in various forms such as distance matrices, MSAs, and dendograms; however, they have different underlying algorithms and output formats. For example, each SNP pipeline uses a different read mapper and SNP caller and might produce a different format to describe their SNP calls. In comparing wgMLST and SNP workflows which are wholly different algorithms, one SNP might be located in an intergenic region, yielding zero allelic differences by wgMLST; on the other hand many SNPs might be located on a single gene, yielding the collapse of multiple SNPs into a single allelic difference.

A reasonable approach to pipeline comparison, therefore, might be at the phylogenetic level. A classic comparison method is the Robinson-Foulds metric, sometimes called the symmetric difference metric, where the number of internal branches that exist in one tree but not the other are counted (Robinson and Foulds, 1981). Another metric is Kuhner-Felsenstein, sometimes called “branch score” which is similar to Robinson-Foulds but calculates the Euclidean distance between each branch’s length (Kuhner and Felsenstein, 1994). Both Robinson-Foulds and Kuhner-Felsenstein metrics are implemented in the Phylip package in the program treedist (Felsenstein, 1989) and in some programming language libraries such as Bio::Phylo (Vos et al., 2011). Both of these classical metrics rely on unrooted trees, and small differences between two trees can artificially magnify the distance between two trees. A more robust tree metric—the Kendall-Colijn—accounts for both tree topology and branch length (Kendall and Colijn, 2015). The Kendall-Colijn metric compares two rooted trees using Euclidean distances from tip to root with a coefficient λ to give more weight to either topology ($\lambda = 0$) or branch length ($\lambda = 1$). One more reasonable approach to pipeline comparison is assessing the distance matrices between two workflows. The Mantel test uses a generalized regression approach to identify correlations between two distance matrices (Smouse et al., 1986). Therefore, if the genome distances from one workflow vs. another workflow are consistently higher but correlate well, the Mantel test will yield a high correlation coefficient.

In this article, we describe the Lyve-SET workflow, demonstrate how it can aid in bacterial foodborne outbreak investigations, and propose methods of comparison with other phylogenetic workflows.

MATERIALS AND METHODS

Implementation

Lyve-SET is a high quality SNP (hqSNP) pipeline, designed to remove lower-quality SNPs from its analysis and increase phylogenetic signal. Lyve-SET has its origins in the original SET algorithm described in Katz et al. (2013). Major changes

in Lyve-SET compared to SET include integrated read cleaning and phage masking, the use of VarScan instead of FreeBayes for SNP calling, improved production of intermediate files in standard formats, and the use of RAxML v8 to infer trees instead of PhyML (Guindon et al., 2010; Garrison and Marth, 2012; Koboldt et al., 2012; Stamatakis, 2014). The source code is available at <https://github.com/lkatz/Lyve-SET> (v1.1.4f, doi: 10.5281/zenodo.163647).

With the default workflow, there is a well-defined audit trail such that it is clear how Lyve-SET was initialized (Table 1) and from where each analysis was derived (i.e., intermediate files are saved). Lyve-SET requires as input a set of raw reads and a phylogenetically related reference genome assembly. Lyve-SET has only been tested with Illumina reads and default settings are optimized for Illumina data, but it can accept FASTQ files from any platform. These steps are depicted in Figure 1.

Although optional, the first recommended step when running Lyve-SET is pre-processing raw sequencing reads. When using the `--read_cleaner` option, reads are cleaned with CG-Pipeline (Kislyuk et al., 2010). The default Lyve-SET options for the CG-Pipeline read cleaner are `--min_quality 15 --min_avg_quality 20 --bases_to_trim 100` which signifies that each read will be trimmed from the 5' and 3' ends up to 100 bp, until a nucleotide has at least a Phred quality of 15. Then, any read with less than an average quality of 20 will be removed. Accordingly, lower-quality reads are removed, trimmed, and/or corrected. Next, phage genes are discovered in the reference genome using BLASTx against the PHAST database with a custom script `set_findPhages.pl` (Camacho et al., 2009; Zhou et al., 2011). A single transduction within an outbreak can introduce changes in thousands of sites when in reality, it is only a single evolutionary change. For example, this event has been observed in *L. monocytogenes* recovered from Italian cheese products in 2012 (Bergholz et al., 2015). Therefore, phage genes on the reference genome are masked in an optimal Lyve-SET analysis. The masked regions are recorded in a BED-formatted file, and a user can also manually edit this file to exclude any other troublesome regions.

The second step is mapping reads of each genome against the reference assembly by SMALT using the `launch_smalt.pl` script (Ponstingl and Ning, 2010). To achieve high-quality mapping, each read’s match to the reference must be 95% identity or above. The expectation in a single-source outbreak is that there will be very few hqSNPs in a dataset; therefore, this identity threshold should help maintain a high accuracy of read-mappings while removing unrelated and error-prone reads. Additionally, the match must be unambiguous within the reference genome—i.e., it cannot match elsewhere equally well—so that repeat regions are masked. One more filter can be optionally applied to avoid calling SNPs in “cliffs.” A cliff is when the read coverage rises or falls dramatically, possibly due to repeat regions, sequencing anomalies, or other factors. To detect these cliffs, a stand-alone script `set_findCliffs.pl` was developed. This script creates a linear trend line in window sizes of 10 base pairs (bp). If the slope of coverage is > 3 reads per bp or < -3 reads per bp, then the region is masked, and SNPs will not be called at the particular locus in the particular genome.

TABLE 1 | Features of Lyve-SET.

	Description	Lyve-SET	kSNP	RealPhy	SNP-Pipeline	SNVPhyl
Repeat detection	Detection of repeat elements that could confound SNP results	0 ^a	0	0	0	1 ^a
Auto-choose reference or reference-free	Independence of a reference genome or a user-defined reference genome to find SNPs	0	1	1	0	0
Removal of distant genomes	Removal of genomes from analysis when they are greater than a certain threshold of SNPs	0	0	0	1	0
Phage detection	Detection and masking of phages	1	0	0	0	0
Cliff detection	Detection and masking of cliffs	1	0	0	0	0
SNP cluster detection	Detection and masking of clustered SNPs	1	0 ^b	0	1	1
Read cleaning	Cleaning and trimming of raw reads	1	0	0	0	0
BAM file for each individual genome	Standardized BAM files that describe the locations of mapped reads	1	0	1	1	1
VCF file for each individual genome	Standardized VCF files that describe the locations of SNPs and evidence supporting them	1	1	1	1	1
Pooled VCF file	Standardized VCF file that describes the locations of all SNPs for all genomes in a single file. This file is created with the <code>bcftools merge</code> command	1	0	0	1	0
Fasta alignment of all sites	Standardized fasta file of all sites across the reference genome, whether they are invariant or SNP sites	1	0	1	0	1
Fasta alignment of SNPs	Standardized fasta file of SNP sites	1	1	1	1	1
Standardized tree file	File representing the phylogeny in a standardized format, e.g., Newick	1	1	1	0	1
Settings for different species	Does the pipeline have customizable settings for different species? Lyve-SET has customized settings using the <code>--presets</code> flag (Table 2)	1	0	0	0	0
Audit trail: repeatability	Displays the path to the SNP pipeline installation and the exact command to repeat the analysis. Lyve-SET provides the command and all explicit and implicit options	1	0	0	1	1
Automated quality control	Reviews the analysis results and describes low-quality results. This quality control can be a review of the length of the multiple sequence alignment, the number of positions masked in each genome, or simply reviewing something minor like the insert length of each genome. Lyve-SET encompasses this quality control step in <code>set_diagnose.pl</code>	1	0	0	1	1

^aAlthough Lyve-SET does not have repeat detection, it does not allow the short-read mapper to place reads where they map equally well in two locations, i.e., repeat regions. SNVPhyl can perform the same function but also straightforwardly identifies repeat regions in the reference genome.

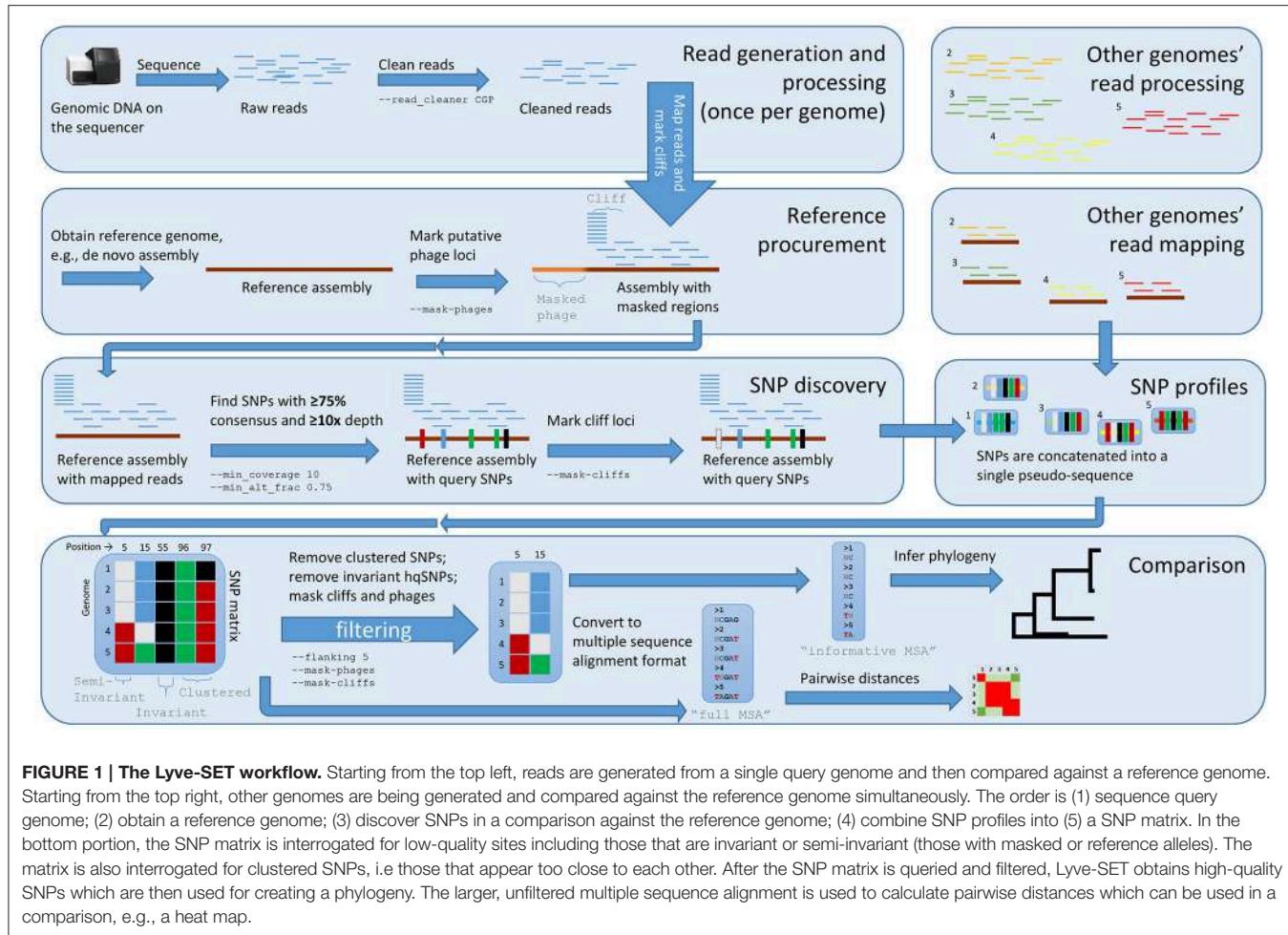
^bAlthough kSNP does not have SNP cluster detection directly, its fundamental algorithm prohibits any SNP from occurring within $k-1$ bp from each other, where k is the length of the kmer. For example on a kmer value of 5, two SNPs must occur at least 4 bp from each other.

Features of Lyve-SET are shown with a comparison of the other SNP pipelines compared in this study. “1” indicates the feature is present; “0” indicates that the feature is absent. A comparison of software-level features, e.g., command-line vs. web interface, has already been performed in Petkau et al. (2016).

The third step is SNP-calling from the read alignments of each genome. Lyve-SET employs the `mpileup2cns` method of VarScan v2.3.7 to find and detect SNPs using the `launch_varscan.pl` script (Koboldt et al., 2012). In this way, VarScan identifies the nucleotide of the query genome at each position of the reference assembly. Any site that has <75% consensus, fewer than 10 reads, or does not have at least two forward and two reverse reads is masked. In the resulting MSA, these masked sites are identified as “N.” Foodborne bacterial pathogens are haploid and so any SNP should be supported by much more than a 50% consensus. Enforcing at least 10 reads at a site helps ensure that a variant is not a random error on a single or few raw reads. Together, these three thresholds increase

support at the SNP-calling stage. However, these values are user-customizable and not hard-coded. In our own investigations, we have modified the settings for many species including *S. enterica*, *E. coli*, and *L. monocytogenes* (**Table 2**). These presets were empirically determined with ongoing outbreaks. Over time, they helped the Lyve-SET results agree with known epidemiology, and so we recorded them. These settings are documented in a configuration file and can be invoked with the Lyve-SET `--presets` option.

The fourth step in Lyve-SET is the creation of a SNP matrix with the `mergeVcf.sh` and `set_processPooledVcf.pl` scripts. Files are merged with the command `bcftools merge` which creates a pooled VCF file. Next, the pooled VCF file



is queried with `bcftools query` to create a tab-delimited matrix consisting entirely of SNPs. If requested, the matrix is also filtered to remove sites with ambiguous nucleotides, invariant sites, and/or clustered SNPs. The resulting matrix or filtered matrix contains only SNPs that pass all filters and therefore contains only hqSNPs. The user may also request annotations for all SNPs if the reference genome is in the GenBank format using SnpEff via the `launch_snpEff.pl` script (Cingolani et al., 2012).

Lyve-SET's fifth step is to convert the SNP matrix to a FASTA-formatted MSA. Using the script `pairwiseDistances.pl` on the MSA, Lyve-SET measures pairwise distances which are helpful in approximating relatedness between taxa. Finally, a phylogeny is inferred using RAxML v8 with the FASTA file containing only hqSNPs, which applies a model for ascertainment bias (Stamatakis, 2014).

All Lyve-SET output files and most intermediate files conform to standardized file formats. Therefore, all results can be viewed in other software if necessary.

Outbreak Clusters

Twelve outbreak clusters of four major foodborne pathogens were queried from the PulseNet database (Table 3; Swaminathan

et al., 2006). PulseNet is a national laboratory network that tracks the subtypes of bacteria causing foodborne illness cases to detect outbreaks. The inclusion or exclusion of isolates for each outbreak was determined using evidence gathered during outbreak investigations including WGS, molecular subtyping, demographic, and exposure data. Isolates from each outbreak were identified; however, some outbreak isolates were excluded when differing from the main outbreak clade by 200 or more hqSNPs. To place these outbreak genomes in a global context, we queried the NCBI *k*-mer trees from April 2016 (Accessions: PDG000000001.428, PDG000000002.629, PDG000000003.184, PDG000000004.427; Data Sheet 1). Each NCBI *k*-mer tree is generated by the NCBI Pathogen Detection Pipeline and is a dendrogram of all publicly available genomes (<https://www.ncbi.nlm.nih.gov/pathogens>). Briefly, NCBI creates high-quality genome assemblies. The MinHash algorithm is applied to each genome, and a Jaccard distance is calculated between each pair of genomes. Then, NCBI creates a tree based on the Jaccard distances. Closely related genomes are refined into subclades using SNPs found among these related assemblies, and a subtree is created with FastME (Lefort et al., 2015). After locating the outbreak clade, we advanced one to three ancestral nodes to acquire a population of potentially related descendant

TABLE 2 | Presets for Lyve-SET.

Name	Settings
lambda	min_coverage=4 min_alt_frac=0.75 mask_phages=0
vibrio_cholerae	min_coverage=10 min_alt_frac=0.75
listeria_monocytogenes	min_coverage=10 min_alt_frac=0.75
salmonella_enterica	min_coverage=20 min_alt_frac=0.95 allowedFlanking=5 mask_phages=1
escherichia_coli	min_coverage=20 min_alt_frac=0.95 allowedFlanking=5 mask_phages=1
clostridium_botulinum	min_coverage=10 min_alt_frac=0.75 allowedFlanking=5 mask_phages=1 mask_cliffs=1

Some settings that have been empirically determined are in a configuration file in the Lyve-SET package. These settings can be revised by individual users in the file *presets.conf*. For many species such as *Campylobacter jejuni*, we have not yet determined the most optimal preset options. However in the future these settings could be added upon or revised following any observations we may make in the due course of outbreak investigations. In each Lyve-SET run, these settings and their values are displayed in the log file, whether or not they were explicitly defined and whether or not the preset configurations were explicitly called.

genomes (**Data Sheet 2**). True positive (TP) isolates are those identified to be associated with the outbreak by PulseNet; true negative (TN) isolates are not associated with the outbreak. For each bioinformatics pipeline to calculate sensitivity (Sn) and specificity (Sp), we also needed to find misidentified genomes, namely the false positives (FP), and false negatives (FN). Sn is calculated as TP/(TP+FN); Sp is calculated as TN/(TN+FP).

Pipeline Parameters

In the following, “out” is the project output directory. The versions of each of these workflows was the most up to date from all stable versions at the time of this work, and default parameters were used unless otherwise specified. All wrapper scripts used for these SNP pipelines can be found at <https://github.com/lkatz/Lyve-SET-paper>.

Lyve-SET v1.1.4f

The `--presets` flag was set according to each taxon (**Table 2**). In this example, the taxon is *listeria_monocytogenes*.

```
launch_set.pl --numcpus 12 --read_cleaner
CGP --presets listeria_monocytogenes out
```

KSNP3 v3.0.0

`Reference_in.txt` contains the reference genome assembly used in the other reference-based methods. Because kSNP is the only SNP pipeline in this study that does not use nucleotide quality scores, the reads were cleaned before running

TABLE 3 | List of outbreaks.

Outbreak code	Species	In outbreak ^a	References
1308MDGX6-1	<i>L. monocytogenes</i>	39, 7, 0	Chen et al., submitted
1408MLGX6-3WGS	<i>L. monocytogenes</i>	19, 64, 1	Jackson et al., 2015; Timme et al., in review
1411MLGX6-1WGS	<i>L. monocytogenes</i>	28, 16, 0	CDC, 2015
1504MLEXH-1	<i>E. coli</i>	17, 2, 0	Tataryn et al., 2014
1405WAEXK-1	<i>E. coli</i>	6, 4, 4	CDC, 2014; Timme et al., in review
1407MNEXD-1	<i>E. coli</i>	6, 10, 1	Health MDo, 2014
1203NYJAP-1	<i>S. enterica</i>	55, 8, 0	Hoffmann et al., 2016; Timme et al., in review
1409MLJN6-1	<i>S. enterica</i>	9, 29, 0	N/A
1410MLJBP-1	<i>S. enterica</i>	5, 10, 0	N/A
0810PADBR-1	<i>C. jejuni</i>	14, 111, 0	Marler-Clark, 2008; Timme et al., in review
1509VTDBR-1	<i>C. jejuni</i>	8, 8, 0	N/A
1602VTDBR-1	<i>C. jejuni</i>	6, 10, 0	N/A

^aThe number of isolates associated with the outbreak, the number of isolates not associated with the outbreak, and the number of isolates with unknown status. Those with unknown status were not used in calculations for tree sensitivity and specificity. Each outbreak is shown with counts of outbreak-associated and non-outbreak-associated isolates.

it. CG-Pipeline was used to clean each read set as shown below, where “uncleaned.fastq.gz” is the original interleaved read set, “sampleDir” is the reads directory used by kSNP, and “cleaned.fastq” is the cleaned interleaved reads. The other specified parameters encode that up to 50 bp were trimmed, the other parameters were auto-picked, and broken pairs were not retained.

```
run_assembly_trimClean.pl -i uncleaned.
fastq.gz -o sampleDir/cleaned.fastq --bases
_to_trim 50 --auto --nosingletons
kSNP3 -k 31 -annotate reference_in.txt
-all_annotations -in in.txt -core -ML -min
_frac 0.75 -CPU $NSLOTS -NJ -vcf -outdir
out
```

RealPhy v112

```
REALPHY_v112 out/samples out/out-readLength
250 -ref reference
```

Snp-Pipeline v0.5.2

```
run.snp_pipeline.sh -c out/snppipeline.conf
-s $scratch_out/samples -m copy -o out
reference.fasta
```

SNVPhyl v1.0

The CLI version of SNVPhyl was run inside of a docker container. Additionally, SNPs were filtered based on density, with the default threshold set to 2 SNPs within a 20 bp window.

```
snvphyl.py --deploy-docker --fastq-dir
fastqs/ --reference-file reference.fasta
--min-coverage 15 --min-mean-mapping
30 --alternative-allele-ratio 0.75
--run-name name --filter-density-window
20 --filter-density-threshold
2 --repeat-minimum-length 150
--repeat-minimum-pid 90 --output-dir out
```

BioNumerics v7.5

wgMLST analysis was performed using tools in the graphical user interface of BioNumerics 7.5 (Applied Maths, Sint-Martens-Latem, Belgium). Briefly, alleles were identified by both an assembly-free k-mer based approach using raw reads and assembly-based BLAST approach based on SPAdes v3.5.0 assembled genomes using the wgMLST *L. monocytogenes* database built in BioNumerics 7.5 (Bankevich et al., 2012). This database contains 4804 loci representing 1748 loci from the Institute Pasteur core scheme (Moura et al., 2016) and 3056 loci representing the pan-genome of *L. monocytogenes* identified from publicly available reference sequences. Once all alleles were assigned to each genome, an unweighted-pair-group-method-with-arithmetic-mean (UPGMA) tree was constructed based on all loci among all the genomes.

Statistical Tests

Three categories of pipeline comparisons were performed: the Sn and Sp of outbreak isolates included in the target outbreak clade, tests of tree topology, and tests of variant positions and distances.

Comparing Trees

If an isolate fell into the same well-supported clade as outbreak isolates (node confidence value $\geq 70\%$), it was counted as a positive. Otherwise, it was counted as a negative. Positives that retrospectively agree with the outbreak investigation were counted as TP; otherwise, FN. The target well-supported clade is defined as a having a confidence value $> 70\%$ and being as complementary as possible for Sn and Sp for each tree. Sn was calculated as $TP/(TP+FN)$. Sp was calculated as $TN/(TN+FP)$. For the following statistical scripts, Perl v5.16.1 and R v3.3.0 were used.

To compare trees, we implemented the Kendall-Colijn and Robinson-Foulds tests (Robinson and Foulds, 1981; Kendall and Colijn, 2015). The Kendall-Colijn test was implemented in the R package Treescape v1.9.17. The background distribution of trees is a set of 10^5 random trees using the APE package in R (Paradis et al., 2004). Each random tree was created with the R function rtree, with the taxon names shuffled. The query tree was compared against the background distribution and then against the Lyve-SET tree. A Z-test was performed; a $p < 0.05$ indicates that the query tree is closely related to the Lyve-SET tree. The Robinson-Foulds metric, also known as the symmetric difference, was implemented in the Perl package Bio::Phylo and was compared against 10^5 random trees generated in BioPerl (Stajich et al., 2002; Vos et al., 2011). The query tree, i.e., an observed tree from wgMLST or from a SNP pipeline, was compared against the random distribution and against the Lyve-SET tree. A Z-test was performed to compare the distances

against the random distribution and the distance vs. Lyve-SET. A significant p -value ($\alpha < 0.05$) indicates that the query tree is more closely related to the Lyve-SET tree topology than would be expected by chance. Given that low-confidence nodes would not be considered during an outbreak investigation, we removed low-confidence nodes (bootstrap support $< 70\%$), potentially creating multifurcating trees, before performing the Kendall-Colijn test. From this transformation, 47 out of 63 trees became multifurcating for this comparison. Only one Lyve-SET tree, the three wgMLST trees, and 12 RealPhy trees remained binary. The Robinson-Foulds test does not tolerate multifurcation; therefore low-confidence nodes were not removed for those tests. Unless otherwise indicated, all trees were midpoint-rooted. All statistical scripts used in this study are available at <https://github.com/lskatz/Lyve-SET-paper>.

Comparing Distances and SNP locations

To compare genetic distances, we plotted each pairwise distance between genomes into a scatter plot, with the x-axis representing Lyve-SET SNPs and the y-axis representing the distance calculated from the other pipeline. This produced one scatter plot per outbreak dataset. Additionally, we used linear regression analysis on each dataset to create a trend line with a slope indicative of calculated distance per Lyve-SET SNP and an R^2 value indicative of goodness-of-fit. We combined all datasets into graphs of each of the four species in this study. Because Lyve-SET is mainly used for outbreak datasets, we also produced scatter plots which only included outbreak-associated genomes such that we could limit the influence of non-outbreak-associated isolates. Jackson et al. (2016) reported an empirical 50-hqSNP distance between outbreak isolates. We observed similar maximum thresholds for all 12 outbreaks in this study for each species. Some distances in the outbreak-only scatter plots were outliers with a clear separation between < 50 and > 100 on the distance axis (**Data Sheet 3**); therefore in the context of analyses comparing only within outbreak-associated isolates, distances > 100 from non-Lyve-SET pipelines were removed.

We also assessed the correlation between the pairwise distance matrices directly using the Mantel test implemented in the R package Vegan v2.4.0 using the Spearman correlation and 1000 permutations (Mantel, 1967; Oksanen et al., 2007). Each query was compared against Lyve-SET.

To compare SNP positions, the set of SNPs from Lyve-SET was used as reference even though no one pipeline can predict with 100% confidence the correct locations of all SNPs. If a query SNP agreed with the position of the Lyve-SET SNP, it was considered as a TP; if the pipeline excluded a position as a SNP that Lyve-SET excluded, then it was a TN. Sn and Sp were calculated as in the test for Sn/Sp of outbreak isolates.

RESULTS

Evaluation of SNP Pipelines Using Outbreak Data Sets

In general, all SNP pipelines in the comparison ascribe outbreak isolates to the outbreak clade with 100% Sn (**Table 4**, **Data Sheet 4**). The one exception is that Snp-Pipeline misclassified a clade of three isolates in the *Salmonella*

TABLE 4 | Summary of 12 pipeline comparisons.

	Lyve-SET	kSNP	RealPhy	Snp-Pipeline	SNVPhyl	wgMLST
Tree sensitivity (Sn) ^a	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Tree specificity (Sp) ^a	100.0%	90.2%	100.0%	100.0%	100.0%	100.0%
Average of Sn and Sp	100.0%	95.1%	100.0%	100.0%	100.0%	100.0%
Kendall-Colijn ($\lambda = 0$) ^b	–	1.26E-02	7.51E-03	9.28E-03	9.15E-02	1.00E-04
Robinson-Foulds ^b	–	3.16E-69	6.79E-40	5.39E-74	9.61E-49	1.55E-147
Mantel	–	0.60	0.77	0.77	0.79	0.74
SNP ratio ^{c,d}	–	0.53, 0.78	0.97, 0.84	1.61, 1.75	0.67, 0.84	0.69, 0.72
Goodness-of-fit (R^2) ^d	–	0.46, 0.42	0.7, 0.75	0.77, 0.3	0.83, 0.68	0.75, 0.72
Genome analyzed ^e	25.9%	0.1%	84.8%	0.3%	82.1%	88.2%

^aAverage percentage from 11 outbreaks. The *S. enterica* outbreak 1203NYJAP-1 was removed as an outlier because all pipelines except wgMLST produced errors with grouping outbreak vs. non-outbreak isolates. Therefore this dataset was removed from the Sn and Sp calculations as an outlier. ^bGeometric mean.

^cNumber of SNPs per Lyve-SET SNP, averaged across 12 outbreaks. For wgMLST, this is the number of alleles per Lyve-SET SNP.

^dThe average for 12 outbreaks. First value is for all data points; second value is for distances between only outbreak-associated genomes.

^eThe average for 12 outbreaks. Percentage of the reference genome included for analysis. For wgMLST, the average percentage was calculated by obtaining each GenBank-formatted file with annotated wgMLST loci and calculating the breadth of coverage for all loci.

More information can be found in **Data Sheets 3, 4**.

1203NYJAP-1 dataset (**Table 3**). Additionally in most instances, the pipelines have 100% Sp as well. Notably for *C. jejuni*, all pipelines yielded 100% Sn and Sp; for *L. monocytogenes* and *E. coli*, all pipelines but one yielded 100%. The *S. enterica* outbreak data caused some difficulty with less-than-perfect Sn and Sp scores for all six pipelines. For four outbreaks associated with *L. monocytogenes*, *E. coli*, and *S. enterica*, kSNP yielded <100% Sp meaning that some isolates not associated with the outbreak were found in the outbreak clade. Additionally the trees of each pipeline were compared against Lyve-SET (**Data Sheets 1, 2**). The Robinson Foulds test reported $p < 0.05$ with the exception of the kSNP trees for outbreaks 1405WAEXK-1 (*E. coli*, $p < 0.625$) and 1410MLJBP-1 (*S. enterica*, $p < 0.674$). However, according to the Kendall-Colijn test for topology (when $\lambda = 0$), at least one tree per pipeline yields a p -value > 0.05 .

The regression analyses show that other SNP pipelines correlate strongly with Lyve-SET (**Figure 2**). The correlation coefficients from RealPhy and Snp-Pipeline are consistently > 0.8 for outbreaks caused by *L. monocytogenes*, *S. enterica*, and *C. jejuni*; SNVPhyl correlates with > 0.8 for *S. enterica*, *E. coli* and *C. jejuni*. Only SNVPhyl has a high correlation with Lyve-SET distances for *E. coli* outbreaks ($R^2 = 0.92$). Overall except for *C. jejuni* outbreaks ($R^2 = 0.89$), kSNP has low correlation with Lyve-SET ($R^2 = 0.69, 0.23, 0.43$). For many of the organism-specific outbreaks tested, viewing a correlation between outbreak-only isolates was difficult because, the range of Lyve-SET SNPs is very low (**Figure S1**). For the *L. monocytogenes* and *E. coli* regression analyses whose Lyve-SET SNPs range 0–43 and 0–16, respectively, only RealPhy and SNVPhyl consistently have a correlation coefficient > 0.8 . In the *S. enterica* and *C. jejuni* analyses whose Lyve-SET distances are small, most distances from other pipelines are also small. However, there are a significant number of data points from kSNP and Snp-Pipeline in the *S. enterica* analysis whose values for Lyve-SET are zero or one, and whose distance values are > 10 . Additionally there are many data points in the *C. jejuni* scatter plot whose

Lyve-SET distances are < 3 and whose Snp-Pipeline distances are > 10 .

Comparison between hqSNP and wgMLST

As a result of the increased utility of wgMLST for outbreak surveillance (Jackson et al., 2016), an important question is how well allelic distances compare with hqSNP distances. The only well-validated wgMLST scheme at the time of this analysis was for *L. monocytogenes*; therefore, hqSNP and wgMLST comparison was performed using the *L. monocytogenes* data sets (**Table 5**; Moura et al., 2016). For pairwise distances found in all isolates (**Figure 3**, panel 1), the correlation coefficient is 0.58. However, when viewing outbreak-only distances (**Figure 3**, panel 3), the correlation coefficient jumps to 0.96. Visually, there are three distinct clusters of pairwise distances for *L. monocytogenes*; therefore, we performed a third regression analysis with Lyve-SET hqSNPs < 255 (**Figure 3**, panel 2). The correlation is highest in this analysis with $R^2 = 0.98$ and a slope of 0.79 allelic differences per Lyve-SET hqSNP.

The discrepancy between large and small distances and their correlations is most likely due to a large variance in the number of hqSNPs per locus. That is to say, if there are many hqSNPs, it is more likely that a single locus contains many hqSNPs and also likely that many loci contain zero or one hqSNP. To test this hypothesis, we counted the number of hqSNPs that intersected each locus in the reference genome of each of the three outbreak datasets. Fourteen percent of all intragenic hqSNPs shared a locus with at least one other hqSNP (**Data Sheet 5**).

DISCUSSION

We built a whole-genome SNP phylogenomics pipeline called Lyve-SET to aid in epidemiological investigations. The design of Lyve-SET was optimized for these investigations. Several features were incorporated to help retain high-quality SNPs, discard low-support SNPs, and generate highly reliable phylogenies (**Table 1**).

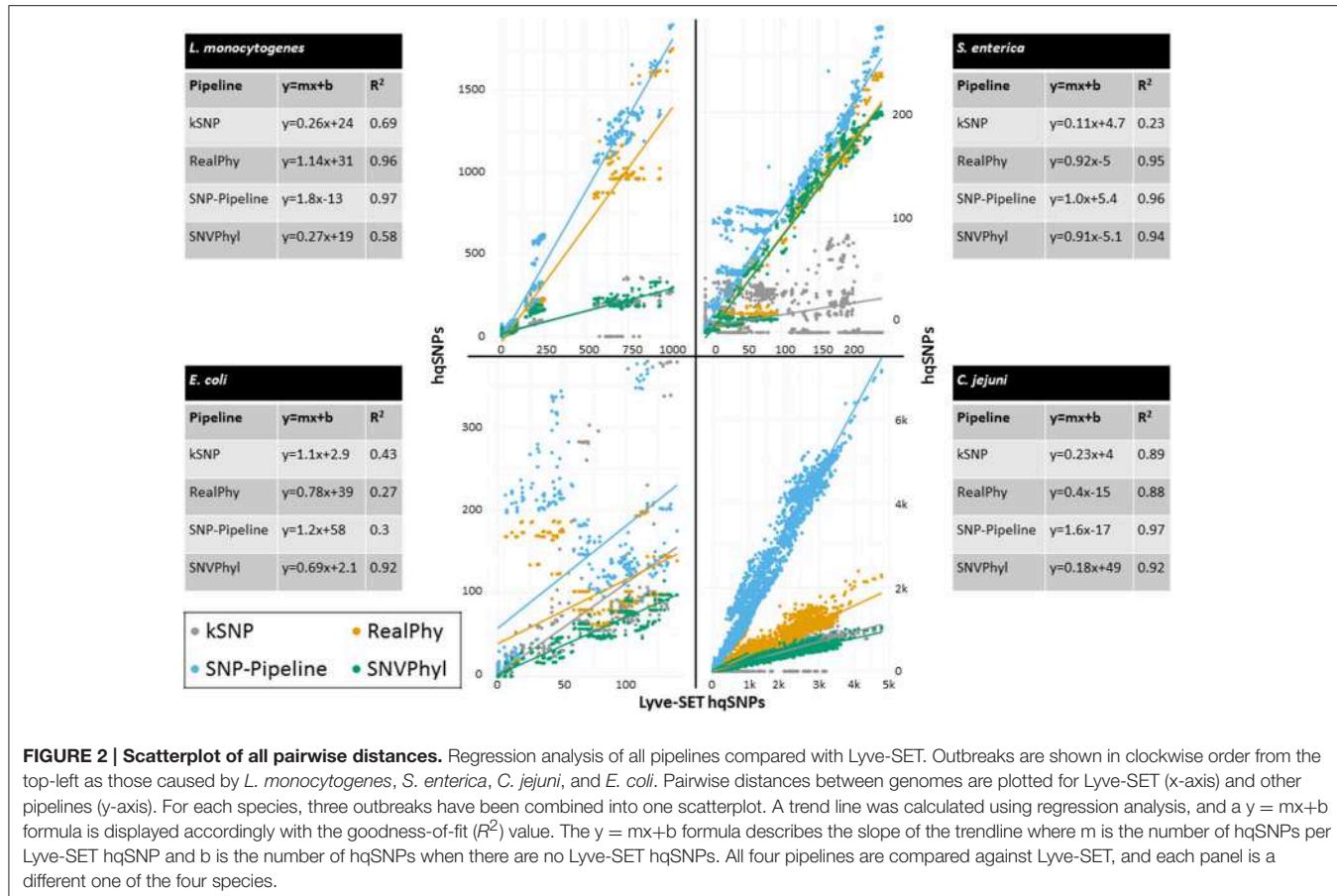


FIGURE 2 | Scatterplot of all pairwise distances. Regression analysis of all pipelines compared with Lyve-SET. Outbreaks are shown in clockwise order from the top-left as those caused by *L. monocytogenes*, *S. enterica*, *C. jejuni*, and *E. coli*. Pairwise distances between genomes are plotted for Lyve-SET (x-axis) and other pipelines (y-axis). For each species, three outbreaks have been combined into one scatterplot. A trend line was calculated using regression analysis, and a $y = mx+b$ formula is displayed accordingly with the goodness-of-fit (R^2) value. The $y = mx+b$ formula describes the slope of the trendline where m is the number of hqSNPs per Lyve-SET hqSNP and b is the number of hqSNPs when there are no Lyve-SET hqSNPs. All four pipelines are compared against Lyve-SET, and each panel is a different one of the four species.

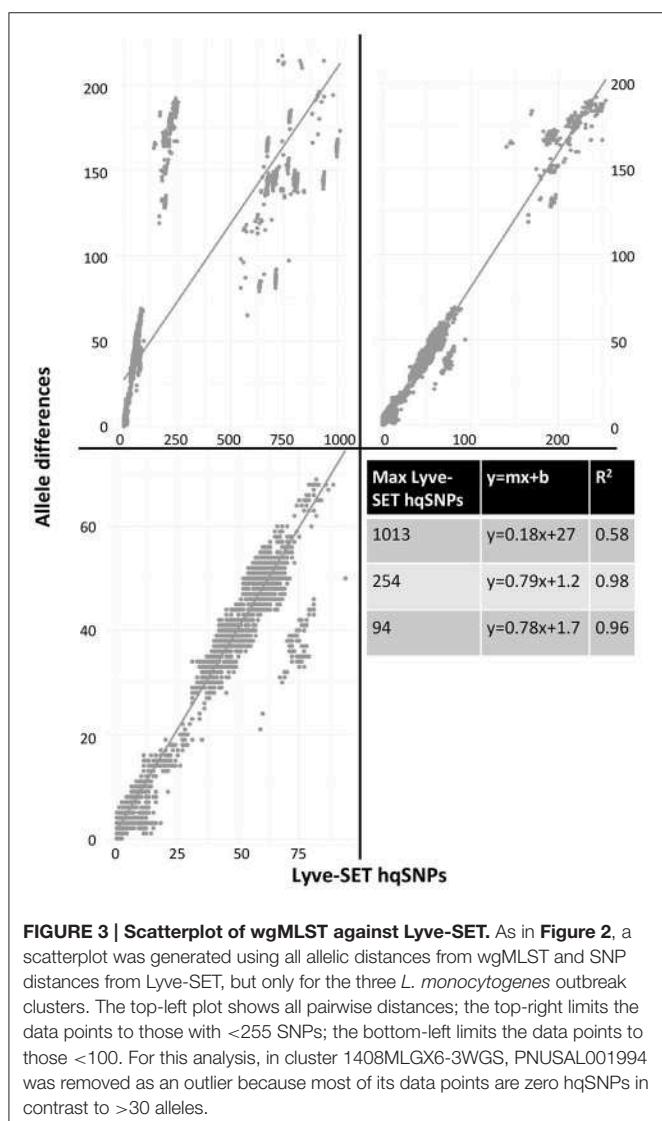
For example, Lyve-SET has the ability to mask “cliffs,” regions where sequencing coverage significantly increases or decreases in a short genomic range. A cliff can be indicative of a repeat region that causes aggregation of short sequencing reads during read mapping. Similarly, other user-defined regions in a BED-formatted file can be supplied to mask unwanted sequences from SNP calling. One example is that, although phages can be useful for typing in their own right (Chen and Knabel, 2008), phage sequences should be removed from a SNP analysis because they often display different rates of mutation than bacterial core genomes. If phages appear to contribute to phylogenetic noise in an investigation, Lyve-SET can provide phage sequence identification with a script `set_findPhages.pl` which is based on a BLAST search against the PHAST database (Zhou et al., 2011). Another way for Lyve-SET to detect troublesome regions is to discard clustered SNPs. For most organisms, this option is preset to 5 bp, such that only one SNP per 5 bp passes the filter. Much like MLST, discarding clustered SNPs reduces noise introduced by horizontal gene transfer. This flanking distance hypothetically should approximate the average recombination cassette length (Vos and Didelot, 2009), but empirically we have found that having a low flanking distance, e.g., 5 bp, is sufficient. There are preset options to customize parameters of each Lyve-SET run for specific organisms (Table 2). For example, a 20x coverage cutoff is used for *S. enterica* while a 10x coverage cutoff

is used for *L. monocytogenes*. In addition to aforementioned features, Lyve-SET, like other SNP pipelines, employs a set of commonly used SNP quality filters such as a percent consensus, a minimum specific coverage threshold, and a requirement of both forward and reverse reads. That is to say, each SNP must be supported by both forward and reverse reads, must have at least a certain number of reads covering each SNP, and must have a certain percentage of reads that agree with the base call. These filters are not only applied to each SNP but also to each position in the genome. Therefore, a SNP should be called for any genome position with a homologous locus in the reference genome, provided that it is not masked and passes all filters. As opposed to most other SNP pipelines, Lyve-SET calls all invariant positions in addition to SNPs in order to perform a rigorous comparison. Therefore in positions where a percentage of genomes have a variant site, all genomes with variant and invariant nucleotide calls can be appropriately compared using various models of evolution. All of these features and filters make Lyve-SET a high-quality SNP pipeline that results in a high-confidence phylogeny, which is often required for outbreak investigations.

Lyve-SET provides a detailed provenance for its outputs including the original Lyve-SET invocation and well-defined intermediate files (Table 1). All intermediate files are in standard file formats (e.g., BAM, VCF) and can be easily inspected with popular third-party tools (Li et al., 2009; Danecek et al., 2011;

TABLE 5 | wgMLST compared against Lyve-SET for outbreaks of *L. monocytogenes*.

	1308MDGX6-1	1408MLGX6-3WGS	1411MLGX6-1WGS
PHYLOGENETIC COMPARISONS			
PKendall-Colijn ($\lambda = 0$)	1E-999	1E-999	1E-4
PRobinson-Foulds	5.53E-108	1.76E-263	3.79E-71
GENOMIC DISTANCE COMPARISONS			
Mantel R^2	0.74	0.73	0.75
Correlation coefficient	0.64	0.73	0.70
Trend line R^2	0.64	0.77	0.84

**FIGURE 3 | Scatterplot of wgMLST against Lyve-SET.** As in Figure 2, a scatterplot was generated using all allelic distances from wgMLST and SNP distances from Lyve-SET, but only for the three *L. monocytogenes* outbreak clusters. The top-left plot shows all pairwise distances; the top-right limits the data points to those with <255 SNPs; the bottom-left limits the data points to those <100. For this analysis, in cluster 1408MLGX6-3WGS, PNUSAL001994 was removed as an outlier because most of its data points are zero hqSNPs in contrast to >30 alleles.

Rodelsperger et al., 2011; Milne et al., 2013; Tamura et al., 2013). In addition to these intermediate files, the output directory has multiple standardized and detailed files (e.g., FASTA, Newick, VCF). These too can be inspected with popular third-party tools

(Danecek et al., 2011; Rodelsperger et al., 2011; Milne et al., 2013; Tamura et al., 2013).

Four other SNP pipelines including kSNP3, RealPhy, Snp-Pipeline, and SNVPhyl were chosen to analyze the same data sets along with Lyve-SET. Each of these pipelines has a history of application to outbreak investigation. In general, all the SNP pipelines evaluated in this study performed well by identifying outbreak isolates in each outbreak clade, yielding >99.5% sensitivity for all pipelines. Except for a few exceptions, all pipelines appropriately excluded non-outbreak-associated isolates. Most notably, the *S. enterica* outbreak 1203NYJAP-1 yielded conflicting results for all pipelines in varying degrees. RealPhy and kSNP identified three and five isolates, respectively, that fit into the outbreak clade for this outbreak. Snp-Pipeline excluded three isolates from the outbreak, reducing its Sn. SNVPhyl produced a star phylogeny, making it difficult to distinguish outbreak from non-outbreak. Lyve-SET included a non-outbreak-associated isolate. The outbreak 1203NYJAP-1 was the sole outbreak that reduced the specificity for Lyve-SET. In all other outbreaks, Lyve-SET correctly classified outbreak vs. non-outbreak isolates 100% of the time.

Due to the increasing utility of wgMLST and the likely co-existence of wgMLST and SNP analyses in surveillance and outbreak investigation of foodborne pathogens, we investigated the concordance of the two methods using three *L. monocytogenes* datasets. By gauging pairwise allelic distances among isolates, we found that the two methods were most consistent with each other when the number of Lyve-SET hqSNPs between any two genomes was <255. The discrepancy between the two methods grew as isolates under study became more divergent. This divergence is most likely due to multiple hits per gene, where one MLST locus could comprise multiple hqSNPs. Therefore if the diversity of the outbreak surpasses the sensitivity of the SNP pipeline and if a wgMLST scheme is available, then a wgMLST approach is more appropriate for an outbreak investigation.

The methodology and datasets reported in this study can help evaluate different pipelines. Due to the non-standard or missing intermediate files of some pipelines and even some output files, it is difficult to compare their results. We recommend that SNP pipelines provide standardized intermediate and output files such as VCF. It is impractical to compare non-standard file formats; fortunately, all pipelines evaluated in this study output a standard Newick tree file and either a FASTA or VCF file which can be used to determine genome distance. We have demonstrated some methods for comparing trees including (1) examining whether individual pipelines could identify outbreak-associated isolates consistent with previous investigation, and (2) whether two trees were significantly similar to each other using the Kendall-Colijn and Robinson-Foulds metrics. We have identified several methods for comparing genome distances. Most notably, the regression analysis has helped identify whether genome distances correlate well between two pipelines and if so, it supplies an equation ($y = mx + b$) that describes how much distance in one pipeline is in another pipeline.

SNP analysis for epidemiological investigations is becoming more common and is a powerful technique (Bertels et al., 2014;

Pettengill et al., 2014; Morganti et al., 2015; Bekal et al., 2016; Jackson et al., 2016; Mercante et al., 2016; Shirey et al., 2016). The methods for analysis for each SNP pipeline have many nuances and are difficult to encompass into a standardized workflow. We have created Lyve-SET to incorporate the many steps of SNP calling into a complete pipeline. Therefore we present Lyve-SET to the community.

AUTHOR CONTRIBUTIONS

Creation of the Lyve-SET software: LK. Selection of datasets: LK. Writing the manuscript: LK, XD, and HC. Comparison of Lyve-SET: LK. Testing Lyve-SET and bug discovery: LK, TG, DW, and HC. Literature search and expertise on phylogenetic and SNP matrix comparisons: AW. Co-authors of the original SET algorithm: AP, CS, GV, and LK. Running and maintaining SNVPhyl: AP, CS, and GV. Running wgMLST through BioNumerics: HC.

FUNDING

This work was made possible through support from the Advanced Molecular Detection (AMD) Initiative at the Centers for Disease Control and Prevention. SNVPhyl development was funded by the Public Health Agency of Canada (PHAC), the Canadian Federal Government Genomics Research and Development Initiative (GRDI) Interdepartmental Shared Priority Project on Food and Water Safety, and Genome BC. The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We would like to thank Kelley Hise, Steven Stroika, Morgan Schroeder, Lavin Joseph, and Eija Trees from PulseNet, and also Amanda Conrad and Kelly Jackson from the Enteric Diseases Epidemiology Branch for helping describe these datasets and uncover details. We would also like to thank the Gen-FS Standards and Analysis working group which created the basis of four of the data sets in this paper (<https://github.com/WGS->

REFERENCES

- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021
- Bekal, S., Berry, C., Reimer, A. R., Van Domselaar, G., Beaudry, G., Fournier, E., et al. (2016). Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent *Salmonella enterica* serovar Heidelberg clone in the context of outbreak investigations. *J. Clin. Microbiol.* 54, 289–295. doi: 10.1128/JCM.02200-15
- Bergholz, T. M., den Bakker, H. C., Katz, L. S., Silk, B. J., Jackson, K. A., Kucerova, Z., et al. (2015). Determination of evolutionary relationships of outbreak-associated *Listeria monocytogenes* strains of serotypes 1/2a and 1/2b by whole-genome sequencing. *Appl. Environ. Microbiol.* 82, 928–938. doi: 10.1128/AEM.02440-15
- Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B., and van Nimwegen, E. (2014). Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Mol. Biol. Evol.* 31, 1077–1088. doi: 10.1093/molbev/msu088
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421
- Carleton, H., and Gerner-Smidt, P. (2016). Whole-genome sequencing is taking over foodborne disease surveillance. *Microbe*, 11, 311–317.
- CDC (2014). *Multistate Outbreak of Shiga toxin-producing Escherichia coli O121 Infections Linked to Raw Clover Sprouts (Final Update) [cited 2016]*. Available online at: <http://www.cdc.gov/ecoli/2014/o121-05-14/index.html>

standards-and-analysis/datasets). Thank you to the expert panel of epidemiologists and lab subject matter experts who originally reviewed each outbreak in the course of their work. We would like to thank Anna Blackstock for advice in statistics. Thank you to Hannes Pouseele from Applied Maths for preliminary conversation regarding cliffs detection. Thank you to Andrew Huang, Chris Gulvik, and Eishita Tyagi for ongoing and future development of Lyve-SET. Lastly, thank you to the original members of Katz et al. (2013) who helped develop SET, the precursor to Lyve-SET.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.00375/full#supplementary-material>

Figure S1 | Scatterplot of all pairwise distances, restricted to outbreak isolates vs. outbreak isolates. The data from **Figure 2** was filtered to only data points that represent only outbreak isolates vs. outbreak isolates. All non-outbreak isolates have been removed. Due to the low numbers of data points and narrow ranges of SNPs, some trend lines are less reliable.

Data Sheet 1 | All NCBI trees. Datasets were originally obtained using PulseNet outbreak codes. However, to identify likely phylogenetic relatives to each outbreak, we used trees from NCBI consisting of all of a single species. There are four trees used in this study. See Section Materials and Methods, Outbreak clusters, for more details.

Data Sheet 2 | All genomes in this study. Genomes are grouped according to the outbreak code. Outbreak or non-outbreak status is shown under “event,” and a suggested reference genome is given. If a genome’s event status is unknown, it is given a value of –1.

Data Sheet 3 | All visual results for tree and SNP comparisons.

Visualizations from **Data Sheet 4** and other comparisons are displayed here.

Data Sheet 4 | All metrics results for tree and SNP comparisons. All metrics that were determined from comparison tests are displayed in this file according to outbreak and pipeline. The last tab shows how **Table 4** was determined.

Data Sheet 5 | hqSNPs per wgMLST locus. MLST-annotated GenBank files were obtained from BioNumerics from each of the three *L. monocytogenes* outbreak datasets. Lyve-SET was re-run using these as reference genome assemblies. The hqSNPs in each sample was compared against the coordinates of the wgMLST loci. The number of hqSNPs per locus was counted per genome, and these counts were aggregated across outbreaks.

- CDC (2015). *Multistate Outbreak of Listeriosis Linked to Commercially Produced, Prepackaged Caramel Apples Made from Bidart Bros. Apples [Updated September 11, 2015]*. Available online at: <http://www.cdc.gov/listeria/outbreaks/caramel-apples-12-14>
- Chen, Y., and Knabel, S. J. (2008). Prophages in *Listeria monocytogenes* contain single-nucleotide polymorphisms that differentiate outbreak clones within epidemic clones. *J. Clin. Microbiol.* 46, 1478–1484. doi: 10.1128/JCM.01873-07
- Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330
- Davis, S., Pettengill, J. B., Luo, Y., Payne, J., Shpuntov, A., Rand, H., et al. (2015). CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Comp. Sci.* 1:e20. doi: 10.7717/peerj-cs.20
- Felsenstein, J. (1989). PHYLIP (Version 3.6) Phylogeny Inference Package. *Cladistics* 5, 164–166.
- Gardner, S. N., Slezak, T., and Hall, B. G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* 31, 2877–2878. doi: 10.1093/bioinformatics/btv271
- Garrison, E., and Marth, G. (2012). Haplotype-Based Variant Detection from Short-Read Sequencing. *arXiv preprint arXiv:12073907*.
- Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010
- Health MD (2014). Minnesota health officials investigating 13 cases of *E. coli* O111 infection [cited 2016]. Available online at: <http://www.health.state.mn.us/news/pressrel/2014/ecoli071414.html>.
- Hoffmann, M., Luo, Y., Monday, S. R., Gonzalez-Escalona, N., Ottesen, A. R., Muruvanda, T., et al. (2016). Tracing origins of the *Salmonella* bareilly strain causing a foodborne outbreak in the United States. *J. Infect. Dis.* 213, 502–508. doi: 10.1093/infdis/jiv297
- Jackson, B. R., Salter, M., Tarr, C., Conrad, A., Harvey, E., Steinbock, L., et al. (2015). Notes from the Field: Listeriosis Associated with Stone Fruit—United States, 2014. Morbidity and mortality weekly report, Public Health Agency of Canada, 282–283.
- Jackson, B. R., Tarr, C., Strain, E., Jackson, K. A., Conrad, A., Carleton, H., et al. (2016). Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. *Clin. Infect. Dis.* 63, 380–386. doi: 10.1093/cid/ciw242
- Katz, L. S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E. S., Turnsek, M. A., et al. (2013). Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* 4:e00398-13. doi: 10.1128/mBio.00398-13
- Kendall, M., and Colijn, C. (2015). A tree metric using structure and length to capture distinct phylogenetic signals. *arXiv preprint arXiv:150705211*.
- Kislyuk, A. O., Katz, L. S., Agrawal, S., Hagen, M. S., Conley, A. B., Jayaraman, P., et al. (2010). A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics* 26, 1819–1826. doi: 10.1093/bioinformatics/btq284
- Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111
- Kuhner, M. K., and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11, 459–468.
- Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* 32, 2798–2800. doi: 10.1093/molbev/msv150
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- MacCannell, D. (2013). Bacterial strain typing. *Clin. Lab. Med.* 33, 629–650. doi: 10.1016/j.cll.2013.03.005
- Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3140–3145. doi: 10.1073/pnas.95.6.3140
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27, 209–220.
- Marler-Clark (2008). *Hendricks' Farm and Dairy Raw Milk 2008* [cited 2016]. Available online at: <http://www.outbreakdatabase.com/details/hendricks-farm-and-dairy-raw-milk-2008/>
- Mercante, J. W., Morrison, S. S., Desai, H. P., Raphael, B. H., and Winchell, J. M. (2016). Genomic analysis reveals novel diversity among the 1976 Philadelphia Legionnaires' Disease outbreak isolates and additional ST36 strains. *PLoS ONE* 11:e0164074. doi: 10.1371/journal.pone.0164074
- Milne, I., Stephen, G., Bayer, M., Cock, P. J., Pritchard, L., Cardle, L., et al. (2013). Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* 14, 193–202. doi: 10.1093/bib/bbs012
- Morganti, M., Scaltriti, E., Cozzolino, P., Bolzon, L., Casadei, G., Pierantoni, M., et al. (2015). Processing-dependent and clonal contamination patterns of *Listeria monocytogenes* in the cured ham food chain revealed by genetic analysis. *Appl. Environ. Microbiol.* 82, 822–831. doi: 10.1128/AEM.03103-15
- Moura, A., Criscuolo, A., Pousselle, H., Maury, M. M., Leclercq, A., Tarr, C., et al. (2016). Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nat. Microbiol.* 2:16185. doi: 10.1038/nmicrobiol.2016.185
- Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J., et al. (2007). *The Vegan Package*. Community ecology package.
- Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412
- Petkau, A., Mabon, P., Sieffert, C., Knox, N., Cabral, J., Iskander, M., et al. (2016). SNVPhyl: A Single Nucleotide Variant phylogenomics pipeline for whole-genome based microbial genomic epidemiology. *Biorxiv*. doi: 10.1101/092940
- Pettengill, J. B., Luo, Y., Davis, S., Chen, Y., Gonzalez-Escalona, N., Ottesen, A., et al. (2014). An evaluation of alternative methods for constructing phylogenies from whole genome sequence data: a case study with *Salmonella*. *PeerJ* 2:e620. doi: 10.7717/peerj.620
- Ponstingl, H., and Ning, Z. (2010). SMALT-a new mapper for DNA sequencing reads. *F1000Research Posters* 1:313.
- Robinson, D. F., and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147. doi: 10.1016/0025-5564(81)90043-2
- Rödelsperger, C., Guo, G., Kolanczyk, M., Pletschacher, A., Kohler, S., Bauer, S., et al. (2011). Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions. *Nucleic Acids Res.* 39, 2492–2502. doi: 10.1093/nar/gkq1081
- Shirey, B., Johnson, S., Maliha, I., Luquez, C., Hill, K., and Maslanka, S. (2016). “Distinguishing outbreaks of botulism using a reference-based SNP analysis and high quality reference genome sequences of *Clostridium botulinum*,” in *11th Annual Sequencing, Finishing, and Analysis in the Future Meeting*. Santa Fe, NM: Los Alamos National Laboratory.
- Smouse, P. E., Long, J. C., and Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool.* 35, 627–632. doi: 10.2307/2413122
- Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12, 1611–1618. doi: 10.1101/gr.361602
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033
- Swaminathan, B., Gerner-Smidt, P., Ng, L. K., Lukinmaa, S., Kam, K. M., Rolando, S., et al. (2006). Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. *Foodborne Pathog. Dis.* 3, 36–50. doi: 10.1089/fpd.2006.3.36
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197
- Tataryn, J., Morton, V., Cutler, J., McDonald, L., Whitfield, Y., Billard, B., et al. (2014). Outbreak of *E. coli* O157: H7 Associated with Lettuce Served

- at Fast Food Chains in the Maritimes and Ontario, Canada, Dec 2012. Canada Communicable Disease Report, Public Health Agency of Canada, 40(S1):2.
- Vos, M., and Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3, 199–208. doi: 10.1038/ismej.2008.93
- Vos, R. A., Caravas, J., Hartmann, K., Jensen, M. A., and Miller, C. (2011). BIO::Phylo-phyloinformatic analysis using perl. *BMC Bioinformatics* 12:63. doi: 10.1186/1471-2105-12-63
- Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Res.* 39, W347–W352. doi: 10.1093/nar/gkr485

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Katz, Griswold, Williams-Newkirk, Wagner, Petkau, Sieffert, Van Domselaar, Deng and Carleton. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

