Hervé Tettelin
Duccio Medini *Editors*

The Pangenome

Diversity, Dynamics and Evolution of Genomes

OPEN

Springer

The Pangenome

Hervé Tettelin • Duccio Medini
Editors

The Pangenome

Diversity, Dynamics and Evolution of Genomes

OPEN

Springer

Editors

Hervé Tettelin
Department of Microbiology and Immunology, Institute for Genome Sciences
University of Maryland School of Medicine
Baltimore, Maryland, USA

Duccio Medini
GSK Vaccines R&D
Siena, Italy

ISBN 978-3-030-38280-3 ISBN 978-3-030-38281-0 (eBook)
https://doi.org/10.1007/978-3-030-38281-0

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Serendipitous discoveries are fascinating events of science inducing, at times, paradigm shifts that give rise to new disciplines *tout-court*.

This is what happened with *pangenomics*: a novel discipline at the intersection of biology, computer science and applied mathematics, whose discovery, development to state of the art and future perspectives are tentatively collected in this book for the first time, 15 years after its inception.

In simple terms, the pangenome concept is the realization that the genetic repertoire of a biological species, i.e. the pool of genetic material present across the organisms of the species, always exceeds each of the individual genomes and can be, in several cases, "unbounded": an *open pangenome*.

This notion was conceived in 2005 as an unexpected, data-driven outcome of the comparative analyses of a few bacterial genomes. This early example of big data in biology—in which a mathematical model, developed to address a practical question in vaccinology, transformed established concepts—opened biology to the unbounded.

Since then, the advent of next-generation sequencing and computational technologies has afforded the generation of pangenomes from thousands of isolates and non-cultured samples of many microbial species, first, and then of eukaryotes encompassing all the kingdoms of life, confirming and extending the original hypothesis beyond the most ambitious expectations.

The first part of the book, *Genomic diversity and the pangenome concept*, opens with a historical account of the original discovery, the observed analogy between genomic sequences and text corpora that allowed the application of mathematical linguistics to the analysis of genomic diversity and the emergence of the pangenome concept in bacteria.

In the second chapter, the reader will find an extensive introduction of the biological species concept with its challenges, the processes associated with the birth and development of a new species and the implications for its pangenome limits.

The following chapter provides a perspective on genome plasticity, pangenome size and functional diversity from the unique point of view of the bacterium itself, followed in the last chapter of the section by a systematic review of the increasingly sophisticated and performant bioinformatic pipelines that have been made available to the scientific community, transforming pangenomics into a commodity tool for the twenty-first century biologist.

The second part, *Evolutionary biology of pangenomes*, aims at making sense of pangenomics through the explanatory perspective of evolution.

As Theodosius Dobzhansky attested half a century ago,[1] *nothing in biology makes sense except in the light of evolution*. Pangenomes are no exception, as the genetic diversity observed in a species is the direct result of the evolutionary interplay between its member organisms and their environment. The effort is facilitated by the significant advances made in the last decade by mathematical modelling, systems theory and computational simulations, in an attempt to clarify the functional mechanisms underpinning diversity generation at the population level, especially in prokaryotes.

The first chapter of this section[2] moves from the dynamic forces that shape pangenome variations, particularly horizontal gene transfer, to discuss the implications for population structures and their ecological significance.

The second chapter analyses the microevolution of bacterial populations by introducing a neutral phylogenetic framework open to the assessment of natural selection and discusses how to reconstruct the microevolutionary history of an entire pangenome. The relationship between pangenomes and selection is further explored in the following chapter, which proposes a stimulating view of pangenomics based on the economic theory of public goods, resulting in the hypothesis that pangenomes are constructed and maintained by niche adaptation. The section closes with a zoom into the alarming public health crisis of antimicrobial resistance, where the authors consider how the pangenome affects the response to antibiotics, the development of resistance and the role of the selective pressures induced by antibiotics and discuss how the pangenome paradigm can foster the development of effective therapies.

The third part, *Pangenomics: an open, evolving discipline*, takes the reader on a journey through applications of pangenome approaches beyond just genes and sequences for prokaryotes and into the realm of eukaryotes. Indeed, as the pangenome concept evolves and genomes from multiple isolates/individuals within virtually all living species become available, it is important to study and challenge the concept beyond the primary genomic sequence and beyond the bacterial world. While most of the pangenome studies published to date focus on genes as the unit,

[1] Theodosius Dobzhansky, The American Biology Teacher, Vol. 35 No. 3, March, 1973; (pp. 125–129) DOI: https://doi.org/10.2307/4444260

[2] Contributed by the brave scholar who once told the late Prof. Stanley Falkow "this is simply because, Stan, you don't understand population biology" [Conference on "Microbial population genomics: sequence, function and diversity", Novartis Vaccines Research Center, Siena (Italy), 17–19 January, 2007].

any sequence (e.g. promoter, intron, intergenic region and mobile element) could be used as the unit to account for the many levels of variation and regulation governing a population, including entire communities occupying a particular niche.

The first chapter of section three provides a vision of how pangenome analyses can be applied to the study of multiple species within a community or microbiome and how outcomes will lead to the characterization of pan-metagenomes across niches or environments. The second chapter describes procedures to infer the biological impact of pangenomic diversity, translating it into functional pathways and their rendition as phenotypes, or panphenomes. The third chapter brings the additional layer of epigenetic regulation into the picture, describing modification processes, methods to detect them and their relationship with the pangenome. Finally, the application of pangenome studies to other kingdoms of life beyond bacteria is a natural extension of the concept. Chapter four provides a detailed overview of eukaryotic genome projects, their genome dynamics and associated pangenome analyses, while the fifth and last chapter of this book compares and contrasts computational strategies that can be implemented towards the characterization of eukaryotic pangenomes.

We hope that this book, thanks to the extraordinary quality of the contributions from each of the authors involved, will provide a broad readership of life scientists with a useful tool for getting acquainted with—or delving deeper into—the pangenome concept and its theoretical foundations, for getting up to speed with the latest technologies and applications of pangenomics, or simply to explore one of the most exciting novelties of twenty-first century biology.

Should pangenomics continue to develop at the current pace, this volume would soon be outdated by the forthcoming developments, killed by its own success.

However, we believe that the elements captured herein—the serendipitous dynamics of the data-driven discovery and the fundamental mindset shift, the understanding of the mechanisms through evolutionary biology, the perspectives and impacts of pangenomics for all kingdoms of life—might remain as a useful reference for the life science community in the years to come.

Baltimore, MD, USA Hervé Tettelin
Siena, Italy Duccio Medini

Acknowledgement

This book was published as an open-access resource thanks to the financial support provided by GSK Vaccines Srl.

Contents

Part I Genomic Diversity and the Pangenome Concept

Part II Evolutionary Biology of Pangenomes

Part III Pangenomics: An Open, Evolving Discipline

About the Editors

Hervé Tettelin Dr. Tettelin is a Professor of Microbiology and Immunology at the University of Maryland School of Medicine, Institute for Genome Sciences. Over the course of his career, Dr. Tettelin developed extensive expertise in microbial genomics, functional genomics, comparative genomics and bioinformatics. He led seminal genome sequencing and analysis projects for many important human bacterial pathogens and related commensals, including the initial genomes of *Streptococcus agalactiae* (group B *Streptococcus*, GBS).

In collaboration with the group of Dr. Rino Rappuoli (GlaxoSmithKline, former Chiron Vaccines and Novartis Vaccines and Diagnostics), Dr. Tettelin pioneered the fields of reverse vaccinology and pangenome analyses. The former makes use of genomics to identify novel protein candidates for vaccine development, which was first applied to *Neisseria meningitidis*; this approach resulted in the recent commercialization of the Bexsero® (4CMenB) vaccine. The latter is the focus of this book.

Dr. Tettelin has conducted many studies of bacterial diversity and transcriptional profiling using DNA microarrays and RNA-seq, as well as functional genomics analyses to identify genes essential for virulence using Tn-seq. He has also supervised the development of bioinformatics tools to compare closely related bacterial genomes in the context of infectious diseases.

Duccio Medini Dr. Duccio Medini is a Data Scientist and Pharma Executive, currently serving as Head of Data Science and Digital Innovation for GSK Vaccines Research and Development. After graduating in Theoretical Physics and receiving his Ph.D. in Biophysics from the University of Perugia, Italy, and the Northeastern University in Boston, MA, Dr. Medini dedicated his activity at solving biological problems that impact human health globally, by extracting knowledge from genomic, epidemiological, preclinical and clinical data with advanced analytics and data-driven computing.

He studied the diversity of bacterial populations leading to the discovery of the pangenome concept, solving the pangenome structure and dynamics of several pathogens; he contributed to the development of the first universal vaccine against serogroup B meningitis and led the Meningococcal Antigen Typing System (MATS) platform worldwide. Recently, he focused on elucidating the mechanisms of action of vaccines and their impact on infectious diseases through complex systems methodologies and initiated a radical, patient-centric redesign of the data models and infrastructure underpinning clinical vaccines research. He has published 40+ scientific articles, book chapters and patents on the population genomics of bacteria and on mathematical modelling of vaccine effects.

Dr. Medini is Full Professor of Molecular Biology and member of international PhD school committees at the Perugia and Turin Universities in Italy, honorary member of the Cuban Immunology Society, Research Fellow of the ISI Foundation, Overseas Fellow of the Royal Society of Medicine and member of the International Society for Computational Biology.

Part I
Genomic Diversity and the Pangenome Concept

The Pangenome: A Data-Driven Discovery in Biology

Duccio Medini, Claudio Donati, Rino Rappuoli, and Hervé Tettelin

Abstract An early example of Big data in biology: how a mathematical model, developed to address a practical question in vaccinology, transformed established concepts, opening biology to the "unbounded."

Keywords Pangenome · Heaps' law · Reverse vaccinology · Group B Streptococcus · Big data · Unbounded diversity

1 The Quest for a *Streptococcus agalactiae* Vaccine

In August of 2000, a collaboration between Rino Rappuoli's team, including Duccio Medini, Claudio Donati, and Antonello Covacci at Chiron Vaccines in Siena, Italy, and Claire Fraser's group, including Hervé Tettelin at the Institute for Genomic Research (TIGR) in Rockville, MD USA, was established to apply their recently pioneered reverse vaccinology approach (Pizza et al. 2000; Tettelin et al. 2000) to the problem of neonatal Group B *Streptococcus* (GBS, or *Streptococcus agalactiae*) infections (Fig. 1a). The collaboration also included Dennis Kasper, Michael Wessels, and colleagues, experts in GBS biology from the Boston Children's Hospital, Harvard Medical School, Boston, MA USA.

GBS is a leading cause of neonatal life-threatening infections, despite the extensive application of antibiotic prophylaxis. Therefore, a vaccine was dearly needed to

D. Medini · R. Rappuoli
GSK Vaccines R&D, Siena, Italy
e-mail: duccio.x.medini@gsk.com

C. Donati
Computational Biology Unit, Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy

H. Tettelin (✉)
Department of Microbiology and Immunology, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA
e-mail: tettelin@som.umaryland.edu

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_1

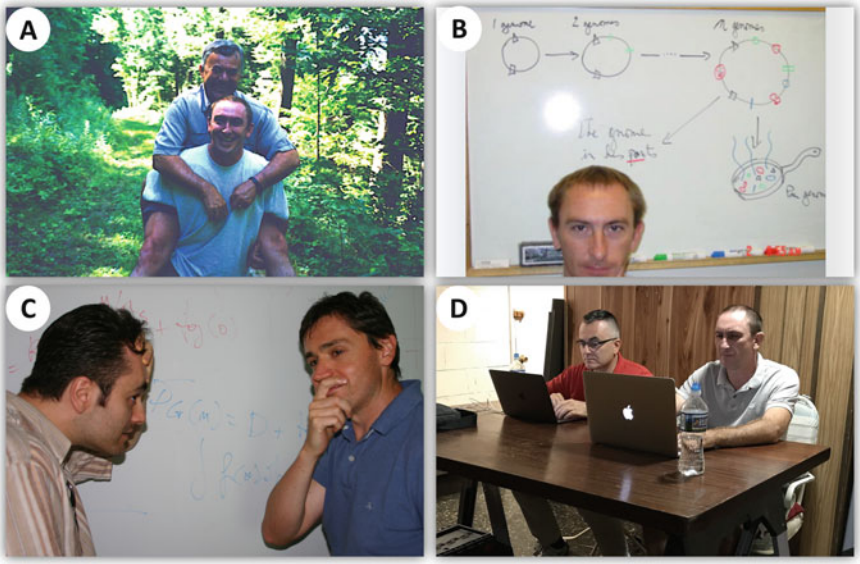


Fig. 1 Pangenome visuals. (**a**) 1999—Plymouth (NH, USA): Rino and Hervé in the woods around the time of initial discussions about the GBS collaboration. (**b**) 2004—Rockville (MD, USA): Pangenome early sketch and (Hervé the) gnome in his pants. (**c**) Early 2005—Siena (Italy): Duccio and Claudio labor over the pangenome formula development. (**d**) 2018—Ellicott City (MD, USA): pangenome book editing, Hervé and Duccio locked in the basement

effectively prevent GBS infections. The manufacturing of a capsular polysaccharide-based vaccine was hindered by the existence and high incidence of at least five different disease-causing serotypes of GBS. Thus, the collaborative team embarked on the development of a GBS protein-based vaccine.

The concept was to use the *Streptococcus agalactiae* genome sequence information to predict proteins likely to be surface exposed and use these in experimental assays for antigenicity and antibody accessibility toward the development of a GBS vaccine via active maternal immunization [for details on GBS reverse vaccinology, see Maione et al. (2005)].

Unlike the case of *Neisseria meningitidis*, with which reverse vaccinology was pioneered right before the GBS project using a single genome, two GBS gap-free genomes were available when the project was initiated, and more genomes were generated early in the course of the project. Indeed, Tettelin et al. [TIGR (Tettelin et al. 2002)] and Glaser et al. [Pasteur Institute, France (Glaser et al. 2002)] independently reported the first two complete gap-free genome sequences of GBS in September of 2002.

At that time, sequencing multiple strains or isolates of the same species was far from commonplace. Both strains, serotype V 2603 V/R and serotype III NEM316, were clinical isolates. Glaser et al. compared their NEM316 genome to that of *Streptococcus pyogenes* (group A *Streptococcus*, GAS) and concluded that 50% of the GBS genes without an ortholog in GAS were located in 14 potential pathogenicity islands enriched in genes related to virulence and mobile elements. Tettelin et al. used a microarray-based comparative genomic hybridization (CGH) approach, whereby they hybridized the genomic

DNA of each of 19 GBS isolates of various serotypes onto a microarray of spotted 2603 V/R gene-specific amplicons, and identified several regions of genomic diversity among GBS isolates, including between isolates of the same serotype (see Fig. 2a).

These separate studies provided the first evidence that a significant amount of genomic information or gene content was variable among closely related streptococcal isolates, challenging the commonly accepted notion that the genome of a single isolate of a given species was sufficient to represent the genomic content of that species. Based on this understanding, the collaborative team decided to generate an additional 6 GBS genomes (Tettelin et al. 2005), selecting isolates from the five major disease-causing serotypes known at the time. The genome of the serotype Ia strain A909 was sequenced to completion in collaboration with the group of Craig Rubens at Children's Hospital and Regional Medical Center, Seattle, WA, USA. The other five strains—515 (serotype Ia), H36B (serotype Ib), 18RS21 (serotype II), COH1 (serotype III), and CJB111 (serotype V)—were sequenced as draft genomes, i.e., no attempt was made to manually close the gaps existing between contigs of the genome assemblies.[1] Comparison of the eight GBS whole-genome sequences confirmed the presence of the regions of genomic diversity previously identified by CGH (see Fig. 2b).

Surprisingly for the time, the shared backbone, or core set of genes present in each of the eight genomes, amounted to only about 80% of any individual genome's gene coding potential. Within these eight genomes, there was no pair that was nearly identical. Instead, each genome contributed a significant number of new strain-specific genes not present in any of the other genomes sequenced. Other sets of genes were shared by some but not all of the genomes.

This large amount of genomic diversity, which was not correlated to GBS serotypes, did not fail to stun members of the investigative team, including the experts in GBS biology. It also prompted an important question that formed the foundation of the pangenome concept: "How many genomes from isolates of the GBS species do we need to sequence to be confident that we identified all of the genes that can be harbored by GBS as a whole?"

This question, motivated by the need to identify all potential vaccine candidates for the species, led to active discussions among the collaborators, the drawing of highly accurate and inspirational scientific sketches (see Fig. 1b), and the decision to develop a mathematical model to determine how many other strains should have been sequenced.

2 When Data Amount and Complexity Exceed What Can Be Done Without Mathematics

The question was clear: "how many genomes...," i.e., the answer had to be a number. And a clear question is always a great way to start.

[1] It should be noted that the COH1 genome, a representative of the highly prevalent disease-causing CC17 clonal complex, was later released as a gap-free genome (NCBI BioProject: PRJEB5232).

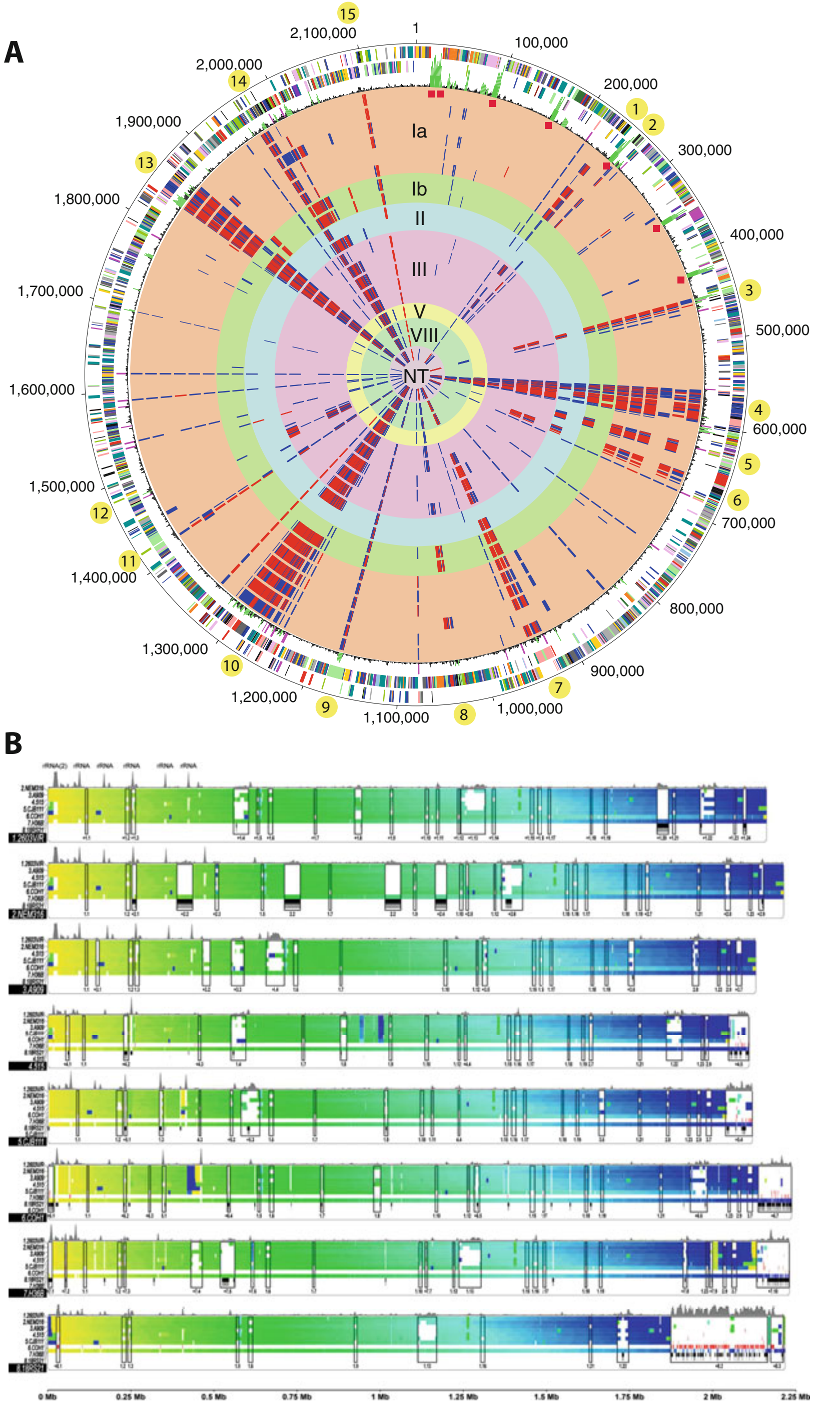


Fig. 2 Group B *Streptococcus* (GBS) genome diversity data that led to the pangenome discovery. (**a**) Comparative genome hybridization (CGH) provided a first hint about the high degree of genomic diversity within the GBS species. This circular representation of the GBS 2603 V/R

When the team in Siena was asked to figure out how to come up with an answer, they were faced with two assumptions, implicit in the question itself. First, the number was expected to be larger than eight, as the presence of specific genes in each of the eight isolates already sequenced suggested. Second, such a finite number was expected to exist.

The whole concept of biological species, a cornerstone of classical cladistics textbooks, had been evolving already toward the "species genome" concept thanks to the genomic revolution. The common knowledge, though, still held a 1:1 relationship between the species and the genome concepts. Consequently, a well-defined genetic repertoire for a bacterial species was the most natural assumption, implying that a finite—and hopefully small—number of genome sequences would be sufficient to exhaust it.

Genomic data had already introduced complexity and size in biology a decade before, when substantial mathematical work had been required to succeed in assembling tens of thousands of Sanger reads into a reconstructed chromosomal sequence (Sutton et al. 1995).

Here complexity and size were growing again, as the population scale of a species was being explored. More mathematical modeling was needed to translate the comparison among genomic data into a number.

Any modeling work starts with arbitrary choices. The first choice—that would remain a cornerstone of pangenome pipelines in the decades to come—was to adopt a reference-free approach.

Population genomics had been explored to that point mostly through cDNA microarrays (CGH), where the experimental design favors the physical comparison of DNA from many isolates with a reference one, usually a well-known laboratory strain used worldwide by the scientific community.

This approach has benefits also for in silico comparative genomics, because the number of comparisons to be performed scales linearly with the number of genome sequences to be compared, i.e., for any new isolate, one more comparison is performed. Also, the high-quality annotation of a well-studied genome can be easily transferred onto the others. However, the reference-based approach introduces strong limitations biasing the comparisons versus one specific individual of the species, which usually has no other ecological merit than having been around in microbiology labs for decades.

Fig. 2 (continued) genome shows predicted ORFs in the two outermost rims and those variable (blue bars) or absent (red bars) in the 19 genomic DNAs hybridized onto the 2603 V/R gene amplicon microarray. Regions of diversity are numbered 1–15 [for details, see (Tettelin et al. 2002)]. (**b**) In silico comparative genomic analysis of 8 GBS genomes confirmed CGH results and revealed additional regions of diversity using each genome as a reference. In this display, genes are arbitrarily color-coded by position in their genome along a gradient from yellow to blue. Genes are then depicted above their ortholog in the reference genome using the color they have in their home genome. Breaks in the color gradient reveal rearrangements and white regions reveal genomic regions absent in query genomes when compared to the reference. Each panel corresponds to each of the eight genomes used as the reference [for details, see Tettelin et al. (2005)]. Copyright 2002, 2005 National Academy of Sciences

Looking for a holistic assessment of a species diversity, the reference-free approach was natural, but it came with the disadvantage of scaling quadratically, i.e., any new genome would have to be compared to all the genomes already considered, leading to significant computational challenges.[2]

The second modeling choice was to use the gene as a unit of comparison or, more precisely, the open reading frames (ORFs) bioinformatically predicted on each genome sequence. Consequently, the analysis focused on an arbitrary subset of the genetic material, ignoring noncoding sequences whose relevance would have been increasingly appreciated in the years to come. Also, it implied accepting a certain number of nucleotide-level polymorphisms as not relevant for the diversity they were trying to model: allelic variants of the same gene would be considered as the same entity, as the problem was not to characterize microevolution—that strains accumulate mutations was well known—but to quantify the amount of "novel" genetic material contributed by each new sequence.

Intuitively, the more genomes analyzed, the fewer new genes (ORFs not observed with sufficient similarity in any other genome) should be identified. To answer the original question ("how many genomes...") the team decided to determine the pace at which new genes would decrease with increasing numbers of genomes sequenced, in order to extrapolate the trend toward the number of genomes corresponding to no new genes identified.

As the number of new genes identified in the n-th genome depends on the selection of both the n-th genome itself and the previous $n - 1$ genomes considered, for each n from 1 to 8 we considered all the $8!/[(n - 1)!\cdot(8 - n)!]$ possible combinations to avoid bias, i.e., a total of 1024 pairwise, whole genome vs. whole-genome comparisons, i.e., ~2 billion gene vs. gene comparisons.

For each n from 1 to 8, we obtained a cloud of values and, following the same approach, the number of core genes (ORFs observed with sufficient similarity in all other genomes) was also measured.

Both new and core gene averages showed the expected decreasing trends, with the number of core genes for GBS decreasing exponentially toward the asymptotic value of 1806.

Surprisingly, though, the decreasing number of new genes was not trending toward zero in any way. Rather, the trend was reasonably reproduced by an exponential decay converging to a fixed value of 33, significantly greater than zero (see Fig. 3a).

In summary, mathematical extrapolation of the trend observed with the first eight genomes indicated that, for every new genome sequenced, new genes would have been discovered, even after a large number of genomes had been sequenced.

The extrapolation had two immediate implications: (i) no number of sequenced genomes would have assured a complete sampling of the GBS species pangenome, because (ii) the genetic repertoire of the species had to be considered as an unbounded entity.

[2] This would have been mitigated a few years later by the introduction of an unbiased, random sampling adjustment (Tettelin et al. 2008).

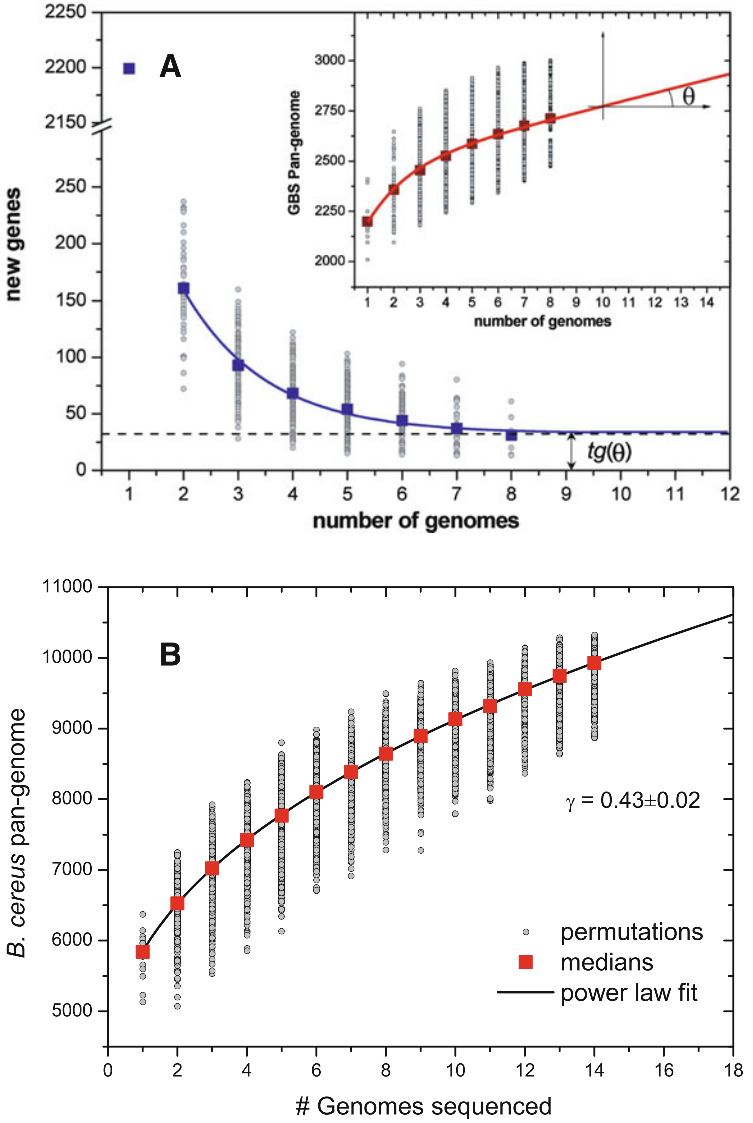


Fig. 3 Mathematical models revealing the "unbounded" pangenome. (**a**) *The first GBS pangenome* (Tettelin et al. 2005), copyright 2005 National Academy of Sciences. The number of specific genes is plotted as a function of the number *n* of strains sequentially added. The blue curve is the least-squares fit of the exponential pangenome function to the data. The extrapolated average number of strain-specific genes is shown as a dashed line. (Inset) Size of the GBS pangenome as a function of *n*. The red curve is the calculated pangenome size with values of the parameters obtained from the fit of the pangenome function to data. (**b**) *The refined power-law pangenome model* (Tettelin et al. 2008). Pangenome of *Bacillus cereus* using medians and a power-law fit. The total number of genes found with the pangenome analysis is shown for increasing numbers of genomes sequenced. Medians of the distributions are indicated by red squares. The curve is a least-squares fit of the power-law pangenome function to medians. The exponent $\gamma > 0$ indicates an open pangenome species

Understandably, the conclusion elicited in the group reactions comprised between complacent irony and the gentle suggestion to redo the work and find the mistake (see Fig. 1c).

So the team did, adding different alignment algorithms, running accurate sensitivity analyses on the thresholds adopted for sequence alignment, applying the same pipeline to other bacterial species known to be less variable as negative control and rechecking every line of the code. Eventually, the team agreed that the extrapolation was correct and novel genes belonging to the GBS species would be found even after sequencing a very large number of genomes. At this point, the team realized the need for a new entity in the genome world to account for those genes that belong to the species but are not present in some genomes. After long discussions, the team agreed on the pangenome concept and described the pangenome of each species by three differentiated components: its core genome, i.e., the genes present in each isolate of the species; its accessory genome, also called initially dispensable, i.e., the genes present in several but not all isolates; and, finally, its strain-specific genes, sampled in one isolate only.

As it would become apparent a few years later, when more genome sequences became available, and for multiple species, a much more accurate description of the trend of new genes would have been provided by a power-law function (derived from the Heaps' law, see Fig. 3b) actually decreasing to zero, as described in more detail in the next section.

But for *S. agalactiae* and some other species, the exponent of the power law was smaller than a critical value, i.e., the decrease of the number of new genes observed with new genomes was so slow, that the size of the pangenome remained an increasing and unbounded function of the number of genomes considered, as is the number of new words discovered in text corpora written in a live language (Heaps 1978). In other words, although the initial modeling work was still incomplete, the conclusion was already correct.

Another critical element that would have gained relevance over time in pangenome analyses, was the heterogeneity of the population sampling. As in any population-modeling exercise, the conclusions at the population scale are heavily dependent on the randomness of the sample, particularly if small, and can be seriously affected by the presence of structure in the population. If only a few, related isolates would be sequenced in an otherwise heterogeneous population, the sample would underestimate the population's diversity. Conversely, if in a population characterized by a few groups of highly similar isolates, we would assess only one genome per group, by extrapolating the measurement to the whole population we would largely overestimate the overall diversity. An effective, albeit incomplete, mitigation of the sampling bias was obtained by replacing the mean of the permutations with medians, which are more robust indicators of centrality.

However, in the original analysis of the eight *S. agalactiae* isolates, one of the more surprising results for the experts of the species' biology, was the lack of any specific relatedness among isolates belonging to the same serotypes, indicating that the phenotypic criteria used to classify the species thus far had no direct relationship with the genomic repertoire of the isolates. From a molecular perspective, this is explained by the fact that genes encoding GBS capsular polysaccharides are part of a single locus, and this locus can be transferred across isolates by lateral gene transfer, showing how the repertoire of dispensable or strain-specific genes can, under specific circumstances, become available to any strain of a given species.

All in all, the answer to the question "How many isolates do we need to sequence to identify all the GBS genes?" was: "there is no such number, the GBS pangenome is open."

The very idea of an unbounded genomic repertoire for a bacterial species was opening the microbiology community to a new way of looking at bacterial species and their anatomy.

While the core-genome remains substantially stable after a few tens of isolates are properly sampled, confirming the genomic consistency of the bacterial species concept, the more isolates are sequenced the more strain-specific genes merge into the accessory genome, expanding the pangenome size.

The underpinning mechanisms and ecological consequences of these dynamics of novelty-generation, spanning the scales of individual mutation, horizontal gene transfer promoted by phage transduction, bacterial conjugation or natural transformation, and population effects would become the object in the years to come of ever-increasing attention of the scientific community (see Fig. 4). A recent example was the observation that the majority of the metabolic innovations in the evolution of *Escherichia coli* arose through the horizontal transfer of single DNA segments (Pang and Lercher 2019).

3 The Vocabulary of Life: Heaps' Law and Pangenomes

In the initial work on *S. agalactiae* (Tettelin et al. 2005), the authors used a decreasing exponential to model the number of new genes discovered in each new genome sequenced. This mathematical function converges asymptotically to a constant value (Fig. 3a and blue curves in Fig. 5). The openness of the pangenome followed from the fact that the best fit of the exponential function to data indicated an asymptotic value significantly higher than zero, i.e., a fixed number of new genes to be discovered in each new genome after the first eight sequenced. Although comforted by the biological diversity observed, such a conclusion was theoretically disturbing because it indicated that, no matter how exhaustively the species would have been sampled, the amount of novelty discovered per new isolate would have remained, on average, constant. A possibility extremely unusual across a wide variety of sampling problems.

In the subsequent work on *H. influenzae* (Hogg et al. 2007), the authors proposed a different approach, focused on the frequency distribution of genes and on the more conservative assumption of a mathematically closed pangenome. However, an increasing number of genomes used to train their model, led to larger predicted size of the pangenome, pointing again toward pangenome openness.

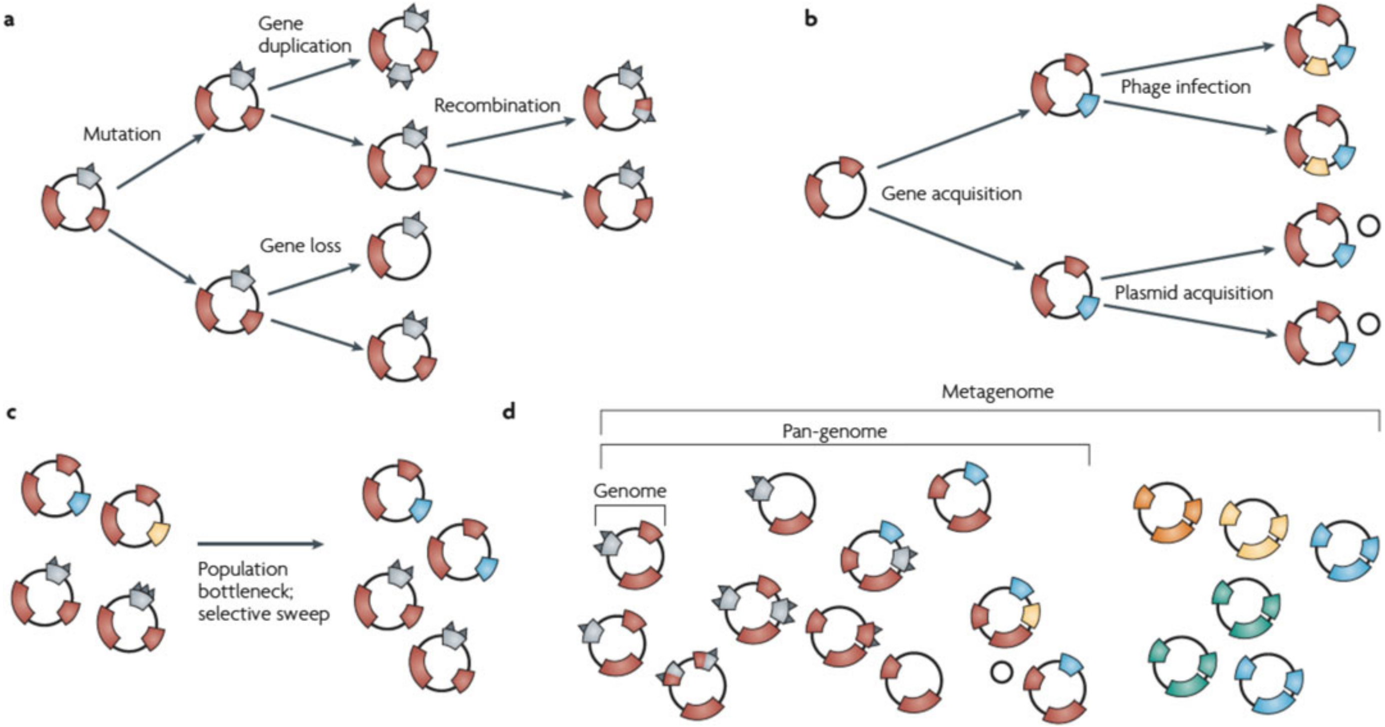


Fig. 4 Molecular evolutionary mechanisms that shape bacterial species diversity: one genome, pangenome, and metagenome (Medini et al. 2008). Intra-species (**a**), inter-species (**b**), and population dynamic (**c**) mechanisms manipulate the genomic diversity of bacterial species. For this reason, one genome sequence is inadequate for describing the complexity of species, genera and their interrelationships. Multiple genome sequences are needed to describe the pangenome, which represents, with the best approximation, the genetic information of a bacterial species. Metagenomics embraces the community as the unit of study and, in a specific environmental niche, defines the metagenome of the whole microbial population (**d**)

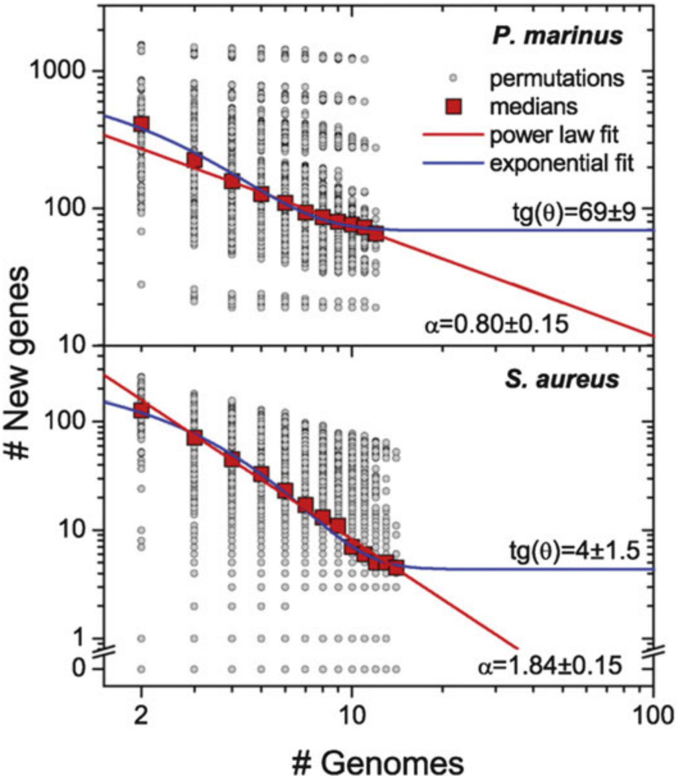


Fig. 5 Power-law regression for new genes (Tettelin et al. 2008). The numbers *n* of new genes are plotted for increasing values of the number *N* of genomes sequenced. Medians of the distributions are indicated by red squares. Blue curves are least-squares fit of the exponential function, as in the original pangenome model. Red curves are least-squares fit of the power-law function. The exponent α determines whether the pangenome is open ($\alpha \leq 1$) or closed ($\alpha > 1$). The top panel shows data for an open pangenome species, *P. marinus*; the bottom panel for a closed pangenome species, *S. aureus*

The collaboration with Ciro Cattuto from the Institute for Scientific Interchange (ISI) Foundation in Turin offered the opportunity to recognize that determining the size of a pangenome was a problem analogous to many similar sampling problems, already addressed when dealing with macroscopic characteristics of complex systems, including human languages.

Before delving into the analogy between genomics and linguistics that allowed to mathematically solve the pangenome problem, a short diversion into the origins of the science of *complex systems* may be useful.

Since the 1970s, a few brilliant minds from disparate academic backgrounds, realized that challenges and opportunities posed by contemporaneity bear a level of complexity exceeding the capacity of established scientific paradigms (Ledford 2015). In 1984, a small group of Nobel laureates and eminent scientists from

Physics, Economics, and other disciplines founded the Santa Fe Institute (Santa Fe Institute) with the visionary ambition of creating a novel science called *complexity* (Waldrop 1993).

That original intuition is at the basis of today's widespread concept of *complex system*, adopted ubiquitously to deal with biological, ecological, economic, technological, and societal systems that cannot be effectively described by linear, inductive approaches, because of the nature of the interactions among system's components, and between the system itself and its environment.

The inductive approach of empirical sciences (i) observes the detailed phenomenology of a system to (ii) infer its underlying dynamics and (iii) uses the inferred laws to describe deterministically the macroscopic properties of the system. For example, (i) observe the movements of planet Earth, Moon and of the Sun to (ii) infer the laws of gravitation and (iii) predict the future trajectory of the planets in the Solar system (Newton 1687).

The approach proposed from the pioneers of complex systems was, in a way, the opposite: (i) start by observing macroscopic, statistical properties shared by multiple systems, (ii) identify a characteristic common to the disparate systems sharing the same property, and (iii) infer generative models, based on that characteristic, capable of accounting for the macroscopic properties observed. For example, (i) observe that in social networks, such as Facebook, few individuals have many connections, and many individuals have few connections, i.e., the frequency "y" of the degree "x" of the network nodes follows a power law "$y = x^{\alpha}$" for some value of the exponent alpha; (ii) confirm that the frequency of words in human languages, of genes in genomes and of inhabitants in cities, all share the same property described for social networks, and all these systems are "modular", i.e., composed of discrete, connected elements; (iii) show that the "preferential attachment" mechanism—according to which the more an element is frequent, the higher the likelihood its frequency will further increase—can be used to generate systems showing the power-law property observed above (Albert and Barabasi 2002).

A similar thinking process led to the solution of the pangenome problem. The rapid accumulation of tens of genome sequences for multiple species had clearly shown that the number of new genes discovered per new genome sequenced follows a decreasing power law, rather than a decreasing exponential trend (see Fig. 5). A similar behavior, for the number of new words discovered upon analyzing increasing numbers of instance texts written in English, had been observed decades earlier by the mathematical linguist Gustav Herdan (1960) and then generalized by Harold Stanley Heaps in the context of information theory as the Heaps' law (Heaps 1978).

When the number of new genes (or words) discovered is a power law of the increasing number of genomes (or text corpora), the overall size of the pangenome (or vocabulary) is also a power law, and the mathematical function depends only on two parameters: the power-law exponent and a proportionality constant. The rate of discovery of new genes is predicted to decrease always toward zero, but the speed of the decrease varies by species. With open pangenomes, such a number is just not decreasing fast enough for the cumulated number of observed genes to level off. Thus, a power-law behavior for the observed number of specific genes allows the possibility of having an open pangenome without requiring that a fixed number of new genes be discovered for each new genome.

In order to complete the approach proposed by the pioneers of complexity, extensive work has been dedicated in recent years to the search for generative models that would account for the macroscopic properties of pangenomes and similar complex systems, including preferential attachment (Albert and Barabasi 2002), self-organized criticality (Bak et al. 1987; Mora and Bialek 2011), and random group formation (Baek et al. 2011). The Heaps' law, however, is only one of such properties displayed by genome data, the other two notable ones being the Zipf's law for the frequency distribution of gene family sizes in complete genomes (Huynen and van Nimwegen 1998) and the "U-shaped" gene frequency distribution (Haegeman and Weitz 2012). The generative models proposed so far could generate some of the observed macroscopic characteristics, but not all at the same time. More recently, a novel mechanism based on a sample space-reducing process (Corominas-Murtra et al. 2015) was proposed, and shown to reproduce naturally the three major properties of pangenomes at once (Mazzolini et al. 2018). Generative processes show how a certain system can be built ("generated") following a pre-defined rule or mechanism; for example, by choosing the elements of the system from an infinite pool of possible components, one after the other randomly. The idea behind the sample space-reducing process for the generation of a certain realization (genome, book) is that when a component (gene, word) is chosen, that choice restricts the space of the possible elements than can be chosen thereafter, permitting only certain other components—but not all—to be added. This assumption seems particularly relevant for genomic and linguistic systems, where the functioning (for genomes) or meaningfulness (for texts) depends on ordered combinations of multiple elements (genes in operons, words in sentences) that are not random (after a restriction enzyme, only a methylation gene produces a restriction-modification system; after a subject, only a verb produces a proposition). For this reason, and considering the relative simplicity of its mathematical implementation, the sample space-reducing process bears promise in the quest for a deeper understanding of the fundamental mechanisms responsible for the generation of pangenomes.

4 Pangenome Vaccinology

The existence of species with an open pangenomes has a profound effect on the selection of potential vaccine candidates identified by a reverse vaccinology approach. Indeed, the accessory genome was found to be an important contributor to protein antigens (Mora et al. 2006) implying that, for many bacterial species, a protein-based universal vaccine would only be possible by including a combination of antigens from the core and the accessory genomes.

The pathogen population structure and dynamics became a key element of vaccine research, paving the way for a modern approach to vaccine discovery known as *pangenomic reverse vaccinology* (Donati et al. 2010; Mora et al. 2006;

Budroni et al. 2011). The key principles of this approach, that expands the reverse vaccinology paradigm based on a single genome sequence (Rappuoli 2000), include reducing bias in isolate selection for genome sequencing (to the extent possible, e.g., carriage vs. invasive isolates, or commensal vs. pathogenic isolates) based on epidemiology, followed by defining the population genomic structure of the species, including its pangenome.

Reverse vaccinology pipelines are then applied to predict the antigenic potential of proteins based on a collection of desired (and undesired features) that they carry, for a recent review on reverse vaccinology pipelines, see Dalsass et al. (2019). Top-ranked vaccine candidate proteins can then be taken through the experimental portion of the vaccine development phase whereby their accessibility and antigenicity are assayed, for instance starting with antigen-based serological typing [for a review on this experimental phase and subsequent phases, please see (Del Tordello et al. 2017)].

It should be noted that the actual transcription, translation, and exposure of a set of selected vaccine protein candidates may vary with the environment, including colonization or infection of various organs, and niches within these organs. The pangenome can inform on these specificities by including isolates with a propensity to target certain organs/niches vs. others (e.g., skin vs. throat isolates of group A *Streptococcus*). Interactions of antigens with host moieties are also key to designing successful candidates.

Ultimately, a combination of pangenomic reverse vaccinology with other multiomics approaches in the context of host–pathogen interactions will better inform the rational design of next-generation vaccine targets and will lead to the most promising formulations to test in vivo.

5 Discussion

In 1946, even before the very discovery of DNA's structure (Watson and Crick 1953), Joshua Lederberg and Edward L. Tatum had demonstrated the existence of "sexual" genetic recombination in bacteria (Lederberg and Tatum 1946), a discovery that granted them the Nobel Prize in 1958.

Horizontal DNA exchange in bacteria has been the subject of intense research ever since so, by the time the pangenome came along, the concept of diversity in bacterial species had been ingrained in the scientific community for half a century already. However, possibly because of the hubris induced by the breakthrough of first-generation genome sequencing technologies, for more than a decade the community had inadvertently reverted to a "pre-Lederberg-Tatum" mindset, considering each of the few genomes generated at the time as representatives of the respective species' genetic blueprint. Of note, the name *pangenome* came to life after many, long discussions on how to name this new concept possibly reflecting the paradigm-shift that was required, at that time, to recognize the simple evidence of facts.

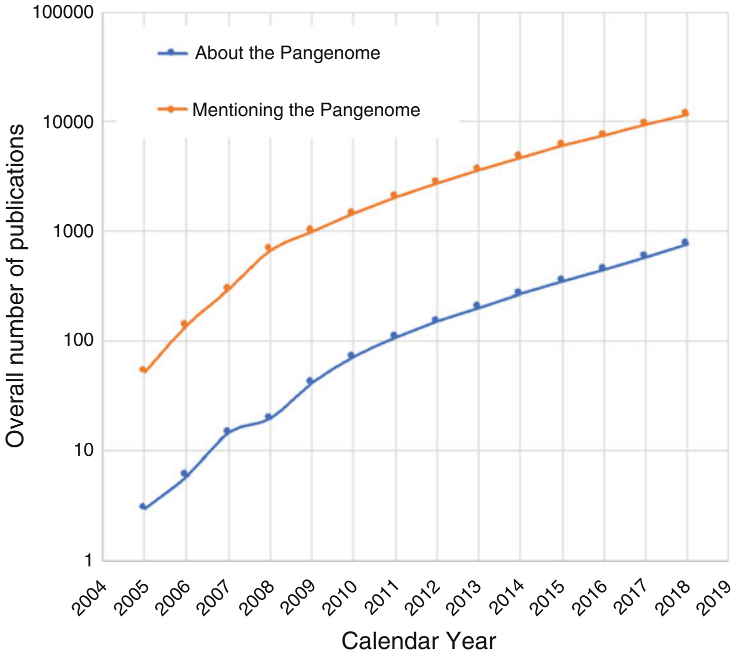


Fig. 6 Pangenome bibliometric data: overall number of pangenome publications since 2005. Blue curve: Number of pangenome publications in PubMed (https://www.ncbi.nlm.nih.gov/pubmed). Query: "pan-genome" [Title/Abstract] OR "pangenome" [Title/Abstract] OR "pan genome" [Title/Abstract]. Orange curve: Number of scientific publications in Google Scholar (https://scholar.google.it) mentioning the pangenome. Query: Exact match to anywhere in the text: "pan-genome" OR "pangenome" OR "pan genome"

The pangenome discovery, at first sight, brought the scientific community back to Earth, to realize that a single genome was far from describing a whole species and that, as a side consequence particularly relevant for the genome pioneers of the time, genome sequencing was there to stay as a flourishing business for decades.

At the same time, though, the pangenome introduced a new dimension in microbiology that could hardly be associated with already established awareness: the concept that the genetic repertoire of a defined biological entity, such as a species, could be unbounded. In a way, the pangenome introduced the infinite in biology, with some humble analogy to what Theodosius Dobzhansky had done 30 years before, much more fundamentally, through the explanatory light of evolution (Dobzhansky 1973).

This could partly explain why, over a relatively short period of time, pangenomics became a discipline in itself (see Fig. 6). From a more practical stance, the impressive acceleration of bacterial genome sequencing was generating high numbers of genes that would not map to species' reference genomes. The new concept offered a conceptual framework to accommodate the wealth of new data, becoming rapidly a must have for any microbial sequencing project.

Thirteen years later (see Fig. 1d), however, two further elements could be identified, that contributed to transforming a specific, empirical question, into a discovery that opened the scientific community to a new research field.

First, a concrete challenge motivated by a burning, unmet medical need, had gathered together people with very different backgrounds, spanning from Biology and Medicine to Physics and Engineering. This collision model, extensively used in modern science and business, promotes ideas that challenge the status quo by facilitating cross-fertilization and lateral thinking. Questioning the serotypes, the team discovered the pangenome. Simple in hindsight but challenging the established paradigm of biological species.

Second, pioneering technological breakthroughs at the bleeding edge, as it was for genome sequencing and assembly at the time, frequently unveils new, unexpected horizons. Not always, though: a critical condition remains the osmotic collaboration between scientists and technology experts mastering the data generation process, to bring in the team awareness of the limitations intrinsic to the data, reducing the risk of hasty misinterpretations, as well as the frustration of missed opportunities.

In conclusion, the pangenome is an early example of mathematical modeling applied to biological Big data: a serendipitous, data-driven discovery from a human health challenge, fostered by technological breakthroughs and people with different backgrounds willing to challenge the status quo. We are deeply grateful to the many investigators worldwide who took the pangenome concept well beyond what could be envisioned at the time, perfected and expanded techniques and applications, and ignited the fascinating evolution of discoveries that the reader now has the opportunity to explore in the remainder of this book.

References

Albert M, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47. https://doi.org/10.1103/RevModPhys.74.47

Baek SK, Bernhardsson S, Minnhagen P (2011) Zipf's law unzipped. New J Phys 13:043004

Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: an explanation of the 1/f noise. Phys Rev Lett 59:381–384

Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV et al (2011) Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci U S A 108(11):4494–4499

Corominas-Murtra B, Hanel R, Thurner S (2015) Understanding scaling through history-dependent processes with collapsing sample space. Proc Natl Acad Sci U S A 112:5348–5353. https://doi.org/10.1073/pnas.1420946112

Dalsass M, Brozzi A, Medini D, Rappuoli R (2019) Comparison of open-source reverse vaccinology programs for bacterial vaccine antigen discovery. Front Immunol 10:113. https://doi.org/10.3389/fimmu.2019.00113

Del Tordello E, Rappuoli R, Delany I (2017) Reverse vaccinology: exploiting genomes for vaccine design. In: Modjarrad K, Koff WC (eds) Human vaccines, emerging technologies in design and development. Academic, Amsterdam, pp 65–86. https://doi.org/10.1016/B978-0-12-802302-0.00002-9

Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. Am Biol Teach 35:125–129

Donati C, Medini D, Rappuoli R (2010) Pangenomic reverse vaccinology. In: Sintchenko V (ed) Infectious disease informatics. Springer, New York, pp 203–221

Glaser P et al (2002) Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. Mol Microbiol 45:1499–1513

Haegeman B, Weitz JS (2012) A neutral theory of genome evolution and the frequency distribution of genes. BMC Genomics 13:196. https://doi.org/10.1186/1471-2164-13-196

Heaps HS (1978) Information retrieval - computational and theoretical aspects. Academic, Orlando, FL

Herdan G (1960) Type-token mathematics. Mouton & Co., The Hague

Hogg JS et al (2007) Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol 8:R103. https://doi.org/10.1186/gb-2007-8-6-r103

Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15:583–589. https://doi.org/10.1093/oxfordjournals.molbev.a025959

Lederberg J, Tatum EL (1946) Gene recombination in *Escherichia coli*. Nature 158:558

Ledford H (2015) How to solve the world's biggest problems. Nature 525:308–311. https://doi.org/10.1038/525308a

Maione D et al (2005) Identification of a universal group B streptococcus vaccine by multiple genome screen. Science 309:148–150

Mazzolini A, Grilli J, De Lazzari E, Osella M, Lagomarsino MC, Gherardi M (2018) Zipf and Heaps laws from dependency structures in component systems. Phys Rev E 98:012315. https://doi.org/10.1103/PhysRevE.98.012315

Medini D et al (2008) Microbiology in the post-genomic era. Nat Rev Microbiol 6:419–430

Mora T, Bialek W (2011) Are biological systems poised at criticality? J Stat Phys 144:268–302

Mora M, Donati C, Medini D, Covacci A, Rappuoli R (2006) Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach. Curr Opin Microbiol 9(5):532–536

Newton I (1687) Philosophiae naturalis principia mathematica. Jussu Societatis Regiæ ac Typis Josephi Streater, Londini. https://www.loc.gov/resource/rbc0001.2013gen20872/

Pang TY, Lercher MJ (2019) Each of 3,323 metabolic innovations in the evolution of *E. coli* arose through the horizontal transfer of a single DNA segment. Proc Natl Acad Sci U S A 116:187–192. https://doi.org/10.1073/pnas.1718997115

Pizza M et al (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 287:1816–1820

Rappuoli R (2000) Reverse vaccinology. Curr Opin Microbiol 3:445–450

Santa Fe Institute Complexity Science. https://www.santafe.edu/about/history

Sutton GG, White O, Adams MD, Kerlavage AR (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci Tech 1:9–20

Tettelin H et al (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 287:1809–1815

Tettelin H et al (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci U S A 99:12391–12396

Tettelin H 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. https://doi.org/10.1073/pnas.0506758102

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11:472–477

Waldrop MM (1993) Complexity: the emerging science at the edge of order and chaos. Simon and Schuster, New York

Watson JD, Crick FH (1953) The structure of DNA. Cold Spring Harb Symp Quant Biol 18:123–131

The Prokaryotic Species Concept and Challenges

Louis-Marie Bobay

Abstract Species constitute the fundamental units of taxonomy and an ideal species definition would embody groups of genetically cohesive organisms reflecting their shared history, traits, and ecology. In contrast to animals and plants, where genetic cohesion can essentially be characterized by sexual compatibility and population structure, building a biologically relevant species definition remains a challenging endeavor in prokaryotes. Indeed, the structure, ecology, and dynamics of microbial populations are still largely enigmatic, and many aspects of prokaryotic genomics deviate from sexual organisms. In this chapter, I present the main concepts and operational definitions commonly used to designate microbial species. I further emphasize how these different concepts accommodate the idiosyncrasies of prokaryotic genomics, in particular, the existence of a core- and a pangenome. Although prokaryote genomics is undoubtedly different from animals and plants, there is growing evidence that gene flow—similar to sexual reproduction—plays a significant role in shaping the genomic cohesiveness of microbial populations, suggesting that, to some extent, a species definition based on the Biological Species Concept is applicable to prokaryotes. Building a satisfying species definition remains to be accomplished, but the integration of genomic data, ecology, and bioinformatics tools has expanded our comprehension of prokaryotic populations and their dynamics.

Keywords Prokaryotes · Speciation · Taxonomy · Biological species concept · Gene flow · Pangenome

L.-M. Bobay (✉)
Department of Biology, University of North Carolina, Greensboro, NC, USA
e-mail: ljbobay@uncg.edu

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_2

1 The Bacterial Species Challenge

Are There Bacterial Species? The taxonomy of microorganisms has been delayed relative to macroscopic organisms, due in part to technical reasons. Evolutionary biologists and population geneticists have originally focused their works on animals and plants, which typically engage in sexual reproduction. For these organisms, speciation mechanisms involve—directly or indirectly—the sustained interruption of gene flow between populations (Dobzhansky 1935; Mayr 1942). The maintenance of gene flow warrants the genetic cohesion of populations, but because prokaryotes do not engage in sexual reproduction stricto sensu, the definition of species has been more elusive in bacteria. It has even been suggested that bacteria cannot and need not be organized into species, but rather represent a series of organisms with different levels of divergence to one another reflecting their past history (Doolittle and Zhaxybayeva 2009; Bapteste et al. 2009). In other words, this view suggests that imposing a grouping of bacteria into species would be purely arbitrary and unreflective of any biologically-relevant process (e.g., cessation of gene flow). However, in practice, microbiologists can usually recognize and designate bacterial isolates based on their different phenotypic characteristics, and comparisons of bacterial genomes indicate that bacteria form clear clusters of highly related individuals, instead of showing a scattered distribution (Riley and Lizotte-Waniewski 2009; Caro-Quintero and Konstantinidis 2012; Konstantinidis et al. 2017), suggesting that they can be organized into species. Ecologically, bacteria can also be identified and clustered based on shared niches and properties (Shapiro and Polz 2014). Altogether, these observations indicate that bacteria can clearly be grouped into genetically and ecologically cohesive entities characteristic of "species", although such species might not be defined based on the same criteria as for sexual organisms. The bacterial species challenge aims to determine the processes that are shaping and maintaining these clusters of cohesive entities.

***Bacterial Genomics and the Case of* Escherichia coli** Before the advent of genotyping methods, microbiologists had to rely exclusively on phenotypic traits to characterize and classify bacteria. Such phenotypic observations offer one criterion for building a species concept, similar to the early approaches used by naturalists to classify animals and plants. However, these early observations showed that it might not be that simple. The seminal work of Oswald Avery and colleagues had strong implications in the field of biology by identifying that DNA—not proteins—was the support of heredity (Avery et al. 1944). But this experiment and previous others further demonstrated that some phenotypic traits could be transmitted horizontally from one bacterial cell to another (Griffith 1928). Although it took several decades to fully understand the extent of horizontal gene transfer in bacteria, this challenging observation contrasted with animals and plants where traits are almost exclusively inherited vertically (i.e., from parent to offspring), indicating that something about bacteria was profoundly different. The development of genetic and genomic techniques further revealed how deeply bacterial genomics differed from animals and plants: related bacteria can differ dramatically in their gene contents and what is typically considered as a bacterial species presents a set of ubiquitous and highly similar genes, the *core-genome*, but also a set of *accessory* genes (also called *dispensable*, *flexible*, or *auxiliary* genes) presenting a scattered distribution (Vernikos et al. 2015). The *pangenome* represents the total gene diversity of a given population: this comprises the total number of distinct orthologs, including core genes and accessory genes (Tettelin et al. 2005; Medini et al. 2005; Vernikos et al. 2015).

The bacteria *Escherichia coli* perfectly illustrates the genomic versatility of prokaryotes. *E. coli* contains approximatively 4400 genes for its model strain K12 MG1655 (Hayashi et al. 2006), but other strains contain up to an additional 1000 genes encoding for a variety of functions (Hayashi et al. 2001). The comparison of only 20 strains of *E. coli* shows that the set of genes shared by all strains—the core-genome—is composed of approximately 2000 genes, but its pangenome approaches readily 18,000 genes (Touchon et al. 2009) and the inclusion of additional strains would necessarily increase this number, as suggested by resampling analyses (Touchon et al. 2009). These numbers indicate that over 50% of the genes of a single strain of *E. coli* consist of accessory genes that do not contain orthologs in the majority of all other strains. Importantly, most of these accessory genes are typically restricted to a single or a small subset of strains, but are often exchanged between strains (Groisman and Ochman 1996; Gogarten et al. 2002; Touchon et al. 2009). Many strains of *E. coli* possess different lifestyles and ecologies broadly ranging from environmental to commensal or pathogenic and these differences can be primarily ascribed to their specific sets of accessory genes (Luo et al. 2011). For example, virulence genes represent a category of extensively studied accessory genes and they appear to be frequently exchanged during *E. coli*'s evolution (Groisman and Ochman 1996; Gogarten et al. 2002).

Although *E. coli* strains present different phenotypes and many different assemblages of accessory genes, they still form a cohesive entity since they share a large number of core genes that are highly similar between all strains of *E. coli* (typically >98% of sequence identity) (Bobay et al. 2013). This situation is problematic for applying phenotype-based classifications in microbiology, as emphasized by the case of *Shigella*. This bacterial "genus" comprises four recognized species (i.e., *S. flexneri*, *S. boydii*, *S. sonnei,* and *S. dysenteriae*), which have been grouped based on shared phenotypic properties (i.e., they are obligate pathogens) (Rolland et al. 1998; Pupo et al. 2000; Escobar-Paramo et al. 2003). However, genomic analyses showed that *Shigella* possesses the same core-genome as *E. coli* with an average of >98% of sequence identity across core genes and core-genome phylogenies revealed that *Shigella* do not form a monophyletic clade (Touchon et al. 2009). What unites *Shigella* together is the presence of shared virulence genes (Buchrieser et al. 2000; Touchon et al. 2009), their serology, and their incapacity to ferment lactose or decarboxylate lysine (Hale and Keusch 1996). In other words, *Shigella* constitutes a subset of *E. coli*'s strains with a shared phenotype conferred by the independent gain of a common set of accessory genes by horizontal gene transfer. It is now recognized that *Shigella* are part of the *E. coli* species, but its taxonomy has not been revised. This example illustrates that the pangenome and its evolutionary dynamics represent a challenge to disentangling the complex relationship between phenotypes, ecology, and genomics in bacteria and how these characteristics correlate with taxonomy.

2 Species Concepts and Operational Definitions

Pragmatic Approaches: Sequence Thresholds One of the goals of a taxonomy is to facilitate communication in the scientific community. To satisfy the need of a coherent microbial taxonomy, pragmatic approaches have been developed in order to define species based on genetic or genomic similarities. Although this does not directly offer insight into how and why a given set of strains constitutes a species, a threshold-based method provides a convenient means to classify strains and revise taxonomy as more comparative genomic data become available. Due to the lack of a theoretical framework of these approaches, such threshold-based methods are often said to define *Operational Taxonomic Units* (OTUs) rather than "species" to emphasize that this is only an operational definition.

Before the rise of the genomic era, species membership was established by shared phenotypic traits and by DNA–DNA hybridization essays, which consist of comparing a newly isolated strain to a reference strain (Brenner et al. 2000) (note that other criteria such as GC content were also considered). The recommended threshold to define species membership was set at 70% of genomic hybridization to the reference strain (Brenner et al. 2000). The emergence of sequencing technologies led to the rise of related approaches. The 16S rRNA subunit has been identified as a universal gene shared by all bacteria and archaea (Woese and Fox 1977) offering the possibility to assess prokaryotic species membership with the same gene marker across all lineages. Analyses revealed that the threshold of 70% identity based on DNA–DNA hybridization assays corresponds approximately to a threshold of 97% identity when using the 16S rRNA subunit (Stackebrandt and Goebel 1994; Ludwig and Klenk 2000; Richter and Rossello-Mora 2009). The use of 16S rRNA thresholds can be applied with ease and allows for the identification of a species by sequencing a single locus. OTU-typing based on the 16S rRNA gene became even more popular with the rise of metagenomic sequencing, where the amplification and sequencing of a fragment of the 16S rRNA gene provides a direct overview of the taxonomic diversity of a given sample without the need of cultivating any of its members. A more recent approach consists of using the entire genome of a strain to calculate the Average Nucleotide Identity (ANI) across all the genes relative to a reference genome of the species (Konstantinidis and Tiedje 2005; Richter and Rossello-Mora 2009). Because protein-coding genes are not as selectively constrained as the 16S rRNA subunit, the ANI threshold used to attain species membership has been empirically defined as 95% based on correlations with 16S sequence threshold used to define species (Konstantinidis and Tiedje 2005; Richter and Rossello-Mora 2009). Considering complete genomes obviously offers a more accurate resolution of sequence divergence.

Sequence thresholds based on single loci or entire genomes present the advantage of defining all prokaryotic species under a standardized framework, but, despite their simplicity, they suffer several technical difficulties. Sequences of the 16S rRNA subunit evolve very slowly and thus sequences from related strains or species typically display little or no informative differences (Kettler et al. 2007). Moreover, multiple copies of the 16S rRNA gene are frequently found in the same genome and they sometimes exhibit different levels of divergence (Acinas et al. 2004). In several cases, the different 16S rRNA copies present in the same genome can display remarkable levels of divergence, such as *Thermoanaerobacter tengcongensis*, which presents 11.6% of sequence divergence between its most different 16S rRNA copies (Acinas et al. 2004). Comparing these sequences would lead to the ironic conclusion that the same bacterial isolate should be classified into two distinct species. A more common criticism against 16S rRNA thresholds is that the divergence of the 16S rRNA gene does not always accurately reflect overall genomic divergence. For instance, the marine bacterium *Prochlorococcus* can be classified as a single species based on 16S rRNA sequences but some strains display only 66% genome-wide identity based on ANI methods (Zhaxybayeva et al. 2009). ANI thresholds are recognized as much more reliable criteria to define species and 16S rRNA alone is of little taxonomic value when complete genome sequences are available (Richter and Rossello-Mora 2009). However, ANI-based methods also suffer inconsistencies. Sequence identity might not be constant along the entire genome (Retchless and Lawrence 2007, 2010) and the identity thresholds used to infer gene orthology can therefore affect the overall ANI value. Perhaps more importantly, ANI metrics are frequently computed against a single reference genome to assess species membership, but the choice of reference genomes is largely arbitrary and historically contingent. In other words, species borders can vary depending on which—or how many—genomes are used as a reference. Finally, using a fixed sequence threshold does not account for the different rates of genomic evolution across phyla (Hugenholtz et al. 2016), which are dictated by parameters like mutation rates, selection coefficients, and effective population sizes (Shapiro 2014) that vary across prokaryotic lineages. Other mechanisms might further lead to differential rates of evolution such as the lack of DNA repair systems (Dorer et al. 2011). Bacterial endosymbionts notoriously evolve at faster rates due to less effective selective pressures imposed by their reduced population sizes (Moran 1996; Moran et al. 2009). As a consequence, the sequence threshold constituting a species in symbiotic bacteria likely corresponds to a different time scale in free-living bacteria (Parks et al. 2018). As a result of all these issues, applying sequence thresholds to define species is convenient but does not anchor a bacterial species concept on a solid theoretical framework.

Phylogenetic Concept Phylogenetic approaches offer another means to classify species. As for sequence thresholds, phylogenetic methods are also a pragmatic approach to define species, although phylogenetic species are defined in the context of evolutionary history (De Queiroz and Gauthier 1994). Besides taking sequence divergence into account, phylogenies typically require species and other taxa to constitute monophyletic groups. Although the concept of monophyly is usually a key feature researched by phylogenetic approaches, it has been argued that *exclusivity* might be preferable over *monophyly* (Velasco 2009; Wright and Baum 2018). Exclusivity is defined as groups of strains/taxa that are more related to one another than other groups without being necessarily monophyletic (Velasco 2009; Wright and Baum 2018). A recent study focusing on *Streptomycetaceae* and *Bacillus* found that exclusive clades can be defined for these taxa, although no objective threshold appears universal (Wright and Baum 2018). An additional and nontrivial advantage of phylogenetic methods is their ability to inform other levels of relationships (e.g., genus and family) and are not restricted to delimiting species. Multiple genome-based phylogenies have been constructed for taxonomic purposes (Garrity 2016; Hugenholtz et al. 2016; Yoon et al. 2017; Parks et al. 2018) and offer a more accurate resolution than 16S rRNA phylogenies (Brochier et al. 2005; Ciccarelli et al. 2006; Thiergart et al. 2014). Akin to sequence thresholds, phylogenetic approaches frequently rely on a single threshold (e.g., a phylogenetic distance) to define species, but recently, a new approach has been developed to reclassify all prokaryotic organisms, while correcting for the uneven evolutionary rates across the tree (Parks et al. 2018). Such approaches offer a universal framework to classify species—and other taxonomic ranks—across the Tree of Life, while correcting for uneven rates of evolution (i.e., defining species with lineage-specific thresholds). The application of these approaches is much more cumbersome than 16S and ANI thresholds, but online tools and resources to place newly sequenced genomes in a reference phylogenetic tree are now available (Parks et al. 2018). The development of such tools and the maintenance of online resources offer the possibility to classify all prokaryotic genomes with ease into a single phylogenetic framework. Although phylogenic methods offer many advantages over sequence threshold methods, they also require comprehensive taxon sampling and can be affected by the underlying phylogenetic model used to reconstruct the tree. Finally, a phylogenetic species concept is still based on ad hoc criteria and does not ambition to identify species based on an explicit speciation model.

The Stable Ecotype Model The stable ecotype model (SEM) is a theoretical framework of bacterial evolution, upon which a microbial species concept can be founded (Cohan 2001; Wiedenbeck and Cohan 2011). In a world without sex, new beneficial alleles can only reach fixation through genome sweep (i.e., fixation of the entire genotype). Therefore, the competition of different bacterial strains for the same resources (the same niche) would lead periodically to the fixation of a single genotype. This model of *periodic selection* implies that most of the diversity of a species is periodically erased, thereby maintaining genetically cohesive entities, i.e., species. Thus, the SEM has the capacity to explain why bacteria form clusters of genomically similar entities. Under this framework, speciation is expected to occur when one strain gains the ability to colonize a different niche (Wiedenbeck and Cohan 2011). By colonizing a different niche, this new population would stop competing against the original population and would not be lost by the periodic selection of a successful genotype of the original population. Note that from the bacterial point of view, a new niche could be as simple as the presence of a new type of carbohydrate and multiple niches are expected to overlap in nature.

A theoretical difficulty of the SEM became apparent when comparing the gene content of bacteria. It became clear that the gene content of a single strain typically represents a very small fraction of the total gene repertoire of the species (i.e., the pangenome) (Tettelin et al. 2005; Medini et al. 2005; Vernikos et al. 2015). This implies that the genetic cohesion of microbial species is only true for a restricted fraction of their genes: their core-genome (Lapierre and Gogarten 2009). The scattered distribution of various accessory genes across strains sharing a highly conserved core-genome cannot be easily reconciled with the SEM. Although a substantial fraction of the pangenome corresponds to mobile elements (Bobay et al. 2013), accessory genes often contribute to the colonization of different niches (Ochman et al. 2000), which implies that the gain and losses of these genes can provide the capacity of a strain to colonize a new niche. This would lead to the disturbing conclusion that a given strain could frequently change species membership by gaining or losing specific sets of accessory genes. Because each genotype virtually contains its own set of accessory genes, each strain could be ascribed to a different ecotype and could be viewed as its own species (Doolittle and Zhaxybayeva 2009; Wiedenbeck and Cohan 2011). This extreme scenario, however, would fail to explain why many bacterial strains present a nearly identical core-genome.

Although the SEM does not easily accommodate the large diversity of accessory genes observed in related bacteria, it has been argued that the definition of an ecotype could be more flexible by encompassing multiple sub-niches (the "nano niche" model) (Wiedenbeck and Cohan 2011). Some strains of a community can acquire alleles or accessory genes specialized in a sub-niche, while remaining part of a broader ecologically-cohesive entity. These specialized strains within an ecotype can be perceived as new species in the making. Nascent speciation might be constantly occurring but need not lead to full speciation (Shapiro and Polz 2014) and this could potentially explain the vast pangenome diversity in bacterial species. Alternative mechanisms have been hypothesized to explain the extensive gene diversity within ecotypes such as a high turnover of accessory genes (Doolittle and Papke 2006) or ecological processes maintaining bacterial diversity such as phage predation ("kill the winner" hypothesis) (Rodriguez-Valera et al. 2009; Thingstad and Lignell 1997) or negative frequency-dependent selection (Cordero and Polz 2014).

While the SEM and related models could provide a coherent explanation of the observation of genomic clusters in the bacterial world—or at least their core-genomes—few results have reported genome sweeps as predicted by the periodic selection expected under the SEM. Multiple studies have overwhelmingly observed that gene sweeps rather than genome sweeps tend to occur under natural conditions (Simmons et al. 2008; Shapiro et al. 2012; Cadillot-Quiroz et al. 2012; Bendall et al. 2016). These results contradict one assumption made by the ecotype model: recombination is negligible relative to selection. Evidence of homologous recombination has been reported for the vast majority of analyzed prokaryotic species (Vos and Didelot 2009; Bobay and Ochman 2017a). That some evidence of homologous recombination exists for most species does not necessarily imply that the rates of homologous recombination are high enough to counteract genome sweeps. A more pertinent metric consists of comparing recombination rate relative to selection: the ratio *r*/*s* (Shapiro and Polz 2014). If selection is overwhelmingly strong relative to recombination, the selected genome is expected to reach fixation before the advantageous alleles are transferred to other genotypes. Because gene sweeps have been more frequently observed than genome sweeps in bacterial species, it seems that the relatively modest levels of homologous recombination in bacteria—in comparison to truly sexual organisms—would suffice to prevent genome sweeps unless extremely beneficial alleles are introduced.

Overall, the accumulation of empirical observations of gene sweeps in natural populations suggest that periodic selection might play a limited role in maintaining genomic cohesion in bacteria. Nevertheless, the SEM remains relevant for effectively clonal species (species with negligible rates of recombination), although the previously cited studies suggest that relatively few species might be effectively clonal (Vos and Didelot 2009; Bendall et al. 2016; Bobay and Ochman 2017a). An inherent difficulty of the SEM and other ecology-based definitions, in general, is the difficulty to gain accurate knowledge on microbial ecology and to identify what objective criteria can be used to define distinct niches. This lack of ecological data appears even more dramatic when compared to the colossal accumulation of genomic data. In the (meta-)genomic era, alternative approaches are needed. Starting from this observation, several authors have suggested the use of a *reverse ecology* approach, where, instead of searching for the genetic variants responsible for ecological segregation, it is more relevant to search for the ecological factors associated with allelic or accessory gene segregation (Shapiro and Polz 2014). The development of a reverse ecology framework potentially offers a powerful tool to extend our comprehension of the ecological factors driving the evolutionary dynamics and the cohesion of bacterial species.

Biological Species Concept Sexual organisms engage in meiotic recombination at each generation and this maintains the genetic cohesion of species (Mayr 1942). The mechanisms leading to speciation in sexual organisms are diverse, can be either pre- or post-zygotic in nature, and are often conceptualized in the context of spatial arrangement of populations (sympatric or allopatric) (Coyne and Orr 2004; De Queiroz 2007). Most models assume that prolonged interruption of gene flow (e.g., zero or few migrants per generation) between two separated populations can lead to the independent accumulation of new alleles and new traits in each population through drift or local adaptation, leading to build up of reproductive incompatibilities and potentially triggering reinforcement, if the two populations are reunited. Other mechanisms, such as the appearance of incompatible alleles or alleles resulting in mating preferences, or even genomic duplications or rearrangements, can also lead to sexual barriers and, therefore, to the interruption of gene flow between populations. While evolution of reproductive barriers is often associated with speciation, it is important to realize that the interruption of gene flow can be either the cause or the consequence of speciation. In all scenarios, however, the interruption of significant gene flow remains associated with speciation, even if the barriers of gene flow can remain somewhat permissive after speciation (Mallet et al. 2007, 2016).

Although bacteria do not engage in true sexual reproduction, it has long been known that they are capable of exchanging DNA (Smith et al. 1993). Because gene flow is a common phenomenon across plants and animals as well as bacteria, this opens the possibility to define bacterial species with the same standards of the biological species concept (BSC) (Dykhuizen and Green 1991; Fraser et al. 2009; Bobay and Ochman 2017a). The fact that bacteria have the capacity to exchange DNA does not necessarily imply that they form biological species; instead, the real challenge is to determine whether the strength of gene flow is sufficient to shape cohesive bacterial units in bacteria, and thus whether common speciation models based on gene flow are applicable to bacteria as well. The question is then: how much and how frequently do they recombine? Can we detect these patterns of gene flow in bacteria as we do for sexual organisms? By "gene flow", I exclusively refer to the replacement of DNA sequences by *homologous recombination* (also referred to as *gene conversion*). Homologous recombination consists of the exchange between two sequences of DNA that typically display a high identity in nucleotide composition (Vulic et al. 1997). In contrast to gene flow, *horizontal gene transfer* (HGT) refers to the gain of new genetic material without the replacement of a homologous sequence. This semantic differentiation allows for the distinction of gene segments of homologous genes that are exchanged (gene flow) versus new genes that are gained (HGT). Note that this distinction permits the differentiation of the outcome of the DNA transfer—homologous replacement or gain of DNA—but it does not necessarily involve different molecular mechanisms since HGT can involve homologous recombination between regions flanking the exchanged sequence (Mell et al. 2011; Croucher et al. 2012; Cordero et al. 2012; Everitt et al. 2014).

Two independent studies have scrutinized a relatively large range of prokaryotic species and came to the conclusion that a small proportion ($<15\%$) of analyzed species do not show substantial signs of gene flow (Vos and Didelot 2009; Bobay and Ochman 2017a). In fact, similar numbers were estimated for viruses and there is growing evidence that the vast majority of cellular and acellular organisms engage in gene flow (Bobay and Ochman 2018a). In addition, many studies have reported that individual loci—rather than entire genotypes—sweep through natural populations (Simmons et al. 2008; Croucher et al. 2011; Shapiro et al. 2012; Cadillot-Quiroz et al. 2012; Bendall et al. 2016; Bao et al. 2016; Porter et al. 2017). These observations imply that gene flow is substantial enough to spread alleles—and even beneficial ones—to the entire population, suggesting the cohesive role of gene flow in bacterial genome dynamics. Importantly, the levels of gene flow across most bacterial species—and their variations—are often substantial enough to be detected using genomic datasets (Bobay and Ochman 2017a). Thanks to the vast accumulation of genomic data, it is possible to identify strains that do not engage in gene flow with the rest of the species (i.e., sexual isolation) by conducting large-scale resampling analyses. This allows to classify sexual eukaryotes, bacteria, archaea, and even viruses under a unique BSC-based species definition.

The delimitation of species based on gene flow is more cumbersome than ANI sequence thresholds, since it requires identification of the core-genome (or a portion thereof) for the tested genome sample and estimation of distances or tree topologies and potentially conducting resampling analyses (Bobay and Ochman 2017b). Similar to phylogenetic methods, it is also possible to compare individual genomes to a database of preprocessed species available online (i.e., ConSpeciFix) (Bobay et al. 2018), which facilitates the classification of newly sequenced data. Detecting and quantifying gene flow remains a delicate endeavor as evidenced by the lack of a consensual methodology to infer homologous recombination. Various methods to estimate recombination rates exist, but they often rely on different models and assumptions regarding the recombination process (Didelot and Falush 2007; Marttinen et al. 2012; Yahara et al. 2014, 2015; Didelot and Wilson 2015; Mostowy et al. 2017), and this contributes to the inference of inconsistent estimates of recombination rates across studies (Bobay et al. 2015). Recently, we introduced a methodology based on the quantification of homoplasies to detect gene flow across large genomic datasets (Bobay and Ochman 2017a; Bobay et al. 2018). Homoplasies are polymorphisms incompatible with vertical inheritance from a shared ancestor and are mostly introduced by gene flow (Bobay and Ochman 2017a). Although the ratio between homoplasic and non-homoplasic polymorphisms does not provide an accurate metric to quantify recombination rates, the detection of homoplasies is rather straightforward and does not rely on complex model assumptions and over parametrization. Interestingly, this homoplasy-based approach appears more robust to genome resampling and gene bootstrapping when compared to ClonalFrameML (Bobay and Ochman 2018b). Inferring gene flow based on homoplasies is limited to the detection of recombination events internal to the dataset and the method does not aim to model imports from external sources. Recombining species can sometimes be misclassified as clonal when multiple sexually isolated genomes are included in the analysis and the sample size is too small to resample and test subpopulations for gene flow; thus, the method is most efficient when large datasets are available and when genetic diversity is high. This limitation will be resolved as more genomes will be sequenced, but, to this date, the analysis of several species can remain inconclusive due to ambiguous signals (Bobay and Ochman 2017a). In addition, the recent accumulation of metagenomic data combined with the development of bioinformatics tools that resolve strain genotypes within metagenomic samples (Nayfach et al. 2016; Pasolli et al. 2017; Truong et al. 2017) constitutes a new source of data readily exploitable to define species based on gene flow.

Because bacteria can sometimes gain genes from other species through HGT, it has been argued that bacteria might not fit a BSC definition in comparison to truly sexual organisms. Species borders are somewhat “fuzzy” for bacteria (Hanage et al. 2005; Hanage 2013) and many studies have detected HGT events in prokaryotes, leading to the conclusion that they might be genomically promiscuous (Popa and Dagan 2011). It should be emphasized, however, that gene flow between species remains very rare when considering the overall time scale of prokaryote evolution, and HGT events occur primarily between related bacteria (Popa et al. 2011). In contrast, gene flow within species is expected to occur at much higher frequencies relative to the acquisition of new genes from external species by HGT (Caro-Quintero et al. 2009; Cadillot-Quiroz et al. 2012; Shapiro et al. 2012; Krause and Whitaker 2015; David et al. 2017). Comparison of ~100 species indicates that most bacteria show clear signs of gene flow and the same method can also retrieve species borders in well classified animals such as humans and *Drosophila* (Bobay and Ochman 2017a). It is well established that sexual eukaryotes are not as well isolated as previously thought (Danchin and Rosso 2012; Syvanen 2012), but introgression and incomplete lineage sorting do not typically prevent defining species borders in truly sexual organisms (Mallet et al. 2016). Although eukaryotic and prokaryotic species borders can be "leaky" and occasionally allow gene flow from external sources, this process need not be prevalent enough to blur species borders (Mallet 2008).

Given the commonality of genomic exchange across diverse types of organisms, a BSC-based definition allows the use of a universal species concept to classify all lifeforms under a biologically relevant definition. What are the implications of applying such a species concept to microbes? Most BSC-species (i.e., bacterial species classified based on the BSC) correspond to closely related genomes that typically present ≥95% ANI (Bobay and Ochman 2017a). However, this is not always true since several BSC-species contain genomes that would not be classified as members of the same species based on ANI thresholds and, conversely, other BSC-species were found to exclude members that would be part of the same species according to ANI thresholds (≥95% ANI) (Bobay and Ochman 2017a). These results are in agreement with analyses showing that a single ANI or phylogenetic threshold fails to define consistent species across prokaryotes (Parks et al. 2018; Wright and Baum 2018). These differences can be putatively ascribed to the use of more-or-less permissive recombination mechanisms across species. Experimental data have suggested that the frequency of homologous recombination decreases exponentially with sequence divergence (Roberts and Cohan 1993; Zawadzki et al. 1995; Vulic et al. 1997; Majewski and Cohan 1998; Majewski et al. 2000) due to the action of the mismatch repair system (Matic et al. 2000). These observations suggest a simple model of sexual isolation in bacteria. The action of the mismatch repair system seems highly variable across taxa (Majewski 2001), which suggests that barriers of gene flow driven by sequence divergence would also be variable across species. In contrast to these observations, there is no systematic negative correlation between recombination and sequence divergence (Bobay and Ochman 2017a) and gene flow has been reported between bacteria presenting relatively divergent genomes (Sheppard et al. 2008; Mell et al. 2011; Cordero et al. 2012), suggesting that sequence divergence plays a limited role in establishing barriers of gene flow. These discrepancies between experimental data and genome analyses can be explained by multiple factors. Firstly, gene flow is detected by the exchange of polymorphisms, and recombination events that do not result in any exchange of polymorphisms can remain invisible to some approaches. This implies that the rates of recombination between highly similar genomes are frequently underestimated. Secondly, selection can potentially have a strong impact in selecting—positively or negatively—alleles exchanged by gene flow, mirroring adaptive introgression or Dobzhansky–Muller incompatibilities in sexual organisms (Mallet et al. 2016). Finally, a simpler explanation might account for these discrepancies. The exponential relationship between sequence identity and recombination rate is based on the observation that nearly identical regions flanking the recombination tract—the minimum efficiently processed segments (MEPS)—are needed to initiate recombination (Shen and Huang 1986; Wiedenbeck and Cohan 2011; Hanage 2016). However, sequence identity need not be high along the entire segment of recombined DNA because recombination requires high sequence identity only along the MEPS, which are only ~26 nt long (Shen and Huang 1986; Wiedenbeck and Cohan 2011; Hanage 2016). This suggests that more variable sequences of DNA might be exchanged as long as a few clusters of nearly identical nucleotides remain available to initiate homologous recombination.

Mixed Model The SEM and a BSC-like model of bacterial evolution need not be fundamentally opposed. A BSC-like model is, by definition, unable to define species borders for clonal species. It is also likely that species with low rates of recombination would appear *effectively* clonal when analyzing genomic data, meaning that the BSC will fail to accurately delimit species in some bacterial groups. For these clades, the SEM appears the most pertinent force maintaining genetic cohesion and therefore is most appropriate to define the borders of these species. The fact that very few studies have reported genome sweeps relative to gene sweeps suggests the prevalence and significance of recombination in bacteria and implies that the vast majority of bacterial species can be defined based on the BSC. Both models could, therefore, be integrated to define species; the SEM for lineages that are effectively clonal and a BSC-like model for species that appear effectively sexual. A key distinction between both models is that the SEM is inherently ecologically centered, whereas a BSC-based model of bacterial evolution does not necessarily involve ecological mechanisms. However, the speciation processes through new niche colonization assumed under the SEM can also lead to speciation under the BSC.

3 Speciation: From Maintenance to Disruption of Genomic Cohesion

Neutral Processes Simulations have provided insightful answers regarding the impact of neutral evolution on the formation of new species. In the absence of recombination, it is expected that some distinct genome clusters would emerge in sympatry (Fraser et al. 2007). However, most of these newly emerged clusters are expected to go extinct through drift. On the other hand, gene flow allows populations to maintain cohesive genomes (Fraser et al. 2007; Friedman et al. 2013). These results suggest that neutral evolution is unlikely to promote the emergence of new species in bacteria, especially in the case of recombining populations. It has been noted that this neutral model of speciation does not consider the potential barrier of gene flow imposed by sequence divergence (Fraser et al. 2007), in which case, it may be possible that divergent genome clusters become more and more sexually isolated. It should be underlined, however, that neutral evolution is expected to drive divergence very slowly, and due to the frequent loss of newly emerged clusters by drift, it is unlikely that population clusters would accumulate enough mutations to impose a substantial barrier of gene flow.

Geography The previous model of neutral speciation has been developed for sympatric populations (i.e., geographically overlapping populations), which is thought to be the preponderant situation in bacteria (Vos 2011; Shapiro and Polz 2015). However, geographic differentiation suggests that allopatric speciation could occur in bacteria (Simmons et al. 2008; Denef et al. 2010; Whitaker et al. 2003; Reno et al. 2009; Krause and Whitaker 2015). Processes resembling allopatric speciation with the interruption of gene flow in bacteriophages targeting different receptors have even been observed in an experimental evolution setting (Meyer et al. 2016). The impact of geography remains elusive since species spanning large continental and oceanic distributions can remain genetically cohesive (Papke et al. 2007; Coleman and Chisholm 2010; Boucher et al. 2011). Recent modeling work has emphasized the impact of niche overlap in bacterial speciation, further revealing the importance of habitat structure in promoting genomic isolation, especially for recombining bacteria (Marttinen and Hanage 2017). The spatial dynamics of microbial distributions remains difficult to characterize and seemingly overlapping populations might not necessarily encounter each other due to fine-scale habitat structure (i.e., mosaic sympatry) (Mallet 2008; Shapiro and Polz 2014).

Recombination Barriers As mentioned above, the initiation of homologous recombination requires the presence of nearly identical short sequences (i.e., MEPS) (Vulic et al. 1997; Majewski and Cohan 1999) and, although relatively divergent sequences can engage in gene flow, sequence divergence can affect recombination rates due to the frequency of available MEPS to initiate recombination. Interestingly, the sequence (MEPS) conservation required to initiate recombination seems to be dependent on the mismatch repair (MMR) system (Matic et al. 2000), which can be more or less permissive across species and strains. The evolution—and sometimes the complete loss—of the MMR system is therefore expected to have a strong impact on sexual isolation in prokaryotes.

Restriction–Modification (RM) systems are frequently used by bacteria to protect themselves against mobile elements and, in particular, bacteriophages (Thomas and Nielsen 2005; Labrie et al. 2010). The presence of different RM systems across strains and species can lead to incompatibilities of gene flow and this has been found to regulate and structure gene flow (Oliveira et al. 2014, 2016). Consequently, the gain or loss of RM systems can have direct consequences on the interruption of gene flow and can potentially lead to speciation. In theory, CRISPR–Cas systems might exhibit similar properties, but since they specifically target a limited number of sequences, they are unlikely to introduce genome-wide incompatibilities. Because of these properties, RM systems can shape the networks of gene flow and the population structure of bacterial species. These systems might drive the establishment of durable barriers of gene flow, potentially leading to speciation.

Gene flow relies on the presence of different vectors and mechanisms capable of disseminating and capturing DNA. The three main mechanisms of DNA transfer, namely transformation, conjugation, and transduction, present diverse degrees of specificity. (i) Transformation does not require cell–to–cell interactions, since environmental DNA is directly taken up by the cell; but recipient cells need to be competent, and relatively few bacteria are known to naturally engage in this process (Johnston et al. 2014). Some bacteria engaging in transformation such as *Neisseria* and *Pasteurellaceae* require the presence of specific DNA uptake sequences or uptake signal sequences (Goodman and Scocca 1988; Scocca et al. 1974; Danner et al. 1982), thereby restricting the range of potential DNA donors to related lineages. Moreover, due to the rapid degradation of DNA when released in the environment this mechanism likely requires close proximity between cells, suggesting that transformation might only mediate gene flow between sympatric populations. (ii) Conjugation involves more constrained transfers of DNA through cell–to–cell contacts, which is mediated by specific pilus interactions and type IV secretion systems (Guglielmini et al. 2013). These conjugative transfers occur primarily between conspecifics, although plasmids have been shown to be occasionally exchanged across much more divergent lineages (Smillie et al. 2010). Because this process requires the direct contact of cells, gene flow mediated by the conjugative apparatus must also occur in sympatry. (iii) Transduction is another route for gene flow where bacterial DNA is packaged within phage particles or gene transfer agents (GTAs) (Lang and Beatty 2007; Popa and Dagan 2011). Phage particles are rarely able to infect multiple species and are often restricted to a subset of strains (Popa et al. 2017). As opposed to transformation and conjugation, phage particles can potentially transport DNA over longer distances (and potentially for long periods of time), suggesting that allopatric—and perhaps anachronistic—populations are able to engage in some levels of gene flow without requiring migration. These three mechanisms, and especially conjugation and transduction, rely on specific molecular signals and are typically restricted to conspecific cells. The overall specificity of these mechanisms is expected to favor gene flow within species rather than between species. Conjugation and transduction also potentially have important consequences for bacterial speciation, since the loss of cell-vector specificity can lead to the partial or complete interruption of gene flow.

Selection As mentioned above, neutral processes are unlikely to lead to bacterial speciation, especially in the case of sympatric recombining populations that co-occur at fine spatial scales (Fraser et al. 2007). This suggests that selection must initiate the formation of distinct genomic clusters, which might eventually lead to selection against genetic intermediates and the cessation of gene flow (Shapiro 2014). Ecological specialization is thought to be a strong force leading to speciation, since the nascent species will present differentially selected EcoSNPs or specialized accessory genes, i.e., alleles or genes specialized in one niche (Shapiro et al. 2012). Simulations have shown that sympatric speciation is more likely when fewer loci are required for speciation and when recombination is reduced (Friedman et al. 2013). As two populations become more and more differentiated, the accumulation of substitutions is expected to reduce gene flow due to epistatic interference (Jain et al. 1999), similarly to Dobzhansky–Muller incompatibilities. Indeed, many loci of the genome coevolve together, and, for instance, central protein complexes such as translation, transcription, and replication complexes require interaction between many central proteins that coevolved together, which could explain why these genes are rarely exchanged by HGT across species, i.e., the "complexity hypothesis" (Jain et al. 1999). Such incompatibilities are expected to be most relevant when populations have significantly diverged and most likely form barriers of gene flow when DNA originates from distant species. However, it is possible that those negatively selected epistatic interactions also contribute to the isolation of more recently diverged populations.

Several studies have demonstrated that the impact of selection on bacterial genome evolution depends on the relative prevalence of selection (s) and recombination rate (r) in sympatric evolution (Shapiro et al. 2009; Friedman et al. 2013; Polz et al. 2013). When selection is much stronger than recombination ($r/s << 1$), the selected allele will lead to the fixation of the entire genotype through genome sweep. The resulting process will be similar to the periodic selection predicted by the SEM. On the other hand, alleles with lower selective coefficients relative to recombination ($r/s >> 1$) are expected to evolve by gene/allele sweep. In this case, selection will be unable to lead to speciation as the selected allele will be exchanged between the population's genotypes by gene sweep. Several studies have attempted to determine whether prokaryotic populations evolve primarily through gene or genome sweeps and, so far, evidence overwhelmingly suggests that gene sweeps are more frequent than genome sweeps (a single case of genome sweep against ~35 cases of gene sweeps (Simmons et al. 2008; Croucher et al. 2011; Shapiro et al. 2012; Cadillot-Quiroz et al. 2012; Bendall et al. 2016; Bao et al. 2016; Porter et al. 2017)). The large prevalence of gene sweeps over genome sweeps is somewhat surprising considering that prokaryotes, as asexual organisms, are thought to display modest rates of gene flow (Wiedenbeck and Cohan 2011). It is, however, difficult to clearly quantify the impact of gene flow on genome evolution (Bobay et al. 2015) and a recent experimental evolution study has shown that gene flow can even lead to the extinction of beneficial alleles (Maddamsetti and Lenski 2018). It is possible that additional factors counteract genome sweeps, such as clonal interference (Lieberman et al. 2014; Maddamsetti et al. 2015) and negative frequency-dependent selection (Cordero and Polz 2014; Takeuchi et al. 2015).

Introgression and HGT from External Species In comparison to the processes acting in sexual organisms, occasional gene flow from external bacteria could be seen as a form of introgression. It has been noted that introgression can sometimes present a source of adaptive alleles in sexual organisms and those transfers can even lead to *hybrid speciation* (Mallet 2007; Rieseberg 1997; Seehausen 2004; Keller et al. 2013). The importance of these processes remains to be explored in prokaryotes. A study comparing the evolution of two *Campylobacter* species—*C. jejuni* and *C. coli*—can be viewed as evidence of bacterial introgression (Sheppard et al. 2008, 2013). Although these results might lead to the complete "despeciation" of the two lineages, it should be noted that the transfer of DNA is asymmetric where one clade of *C. coli* has likely gained alleles from *C. jejuni* but other clades of *C. coli* did not. Interestingly, this case of bacterial introgression appears ecologically-driven based on recent niche overlap (Sheppard et al. 2008). It is, therefore, possible that introgression can result in the same outcomes in prokaryotes, such as hybrid speciation (Shapiro et al. 2016).

Similar to introgression, the gain of new genes from distinct species by HGT offers another means to colonize new niches through ecologically-driven adaptation. The acquisition of antibiotic-resistant genes constitutes a well-documented case, but many other examples have been reported (Ochman et al. 2000; Popa and Dagan 2011). It has been shown that HGT—rather than duplication—plays a predominant role in introducing new paralogs in the pangenome of prokaryotic species (Treangen and Rocha 2011), although these genes frequently come from related species due to genetic incompatibilities (i.e., gene promoters/regulators and codon usage bias) (Sorek et al. 2007; Popa et al. 2017). These acquired genes can mediate the colonization of new niches and can potentially lead to ecology-driven speciation. However, as noted above, accessory genes are not stably associated with a given genotype and tend to be frequently exchanged across strains of a given species (Schubert et al. 2009), indicating that they do not necessarily drive the formation of distinct ecologically specialized entities (Shapiro and Polz 2015).

Summary Across the many forces that can affect speciation, it should be noted that neutral processes such as population dynamics and sequence divergence are unlikely to lead to speciation in bacteria, and that selection seems to be a necessary force by initiating and maintaining speciation. Selection in bacteria can act through two predominant avenues: (i) by driving ecological adaptation to different niches following, for instance, the gain of new genetic material and (ii) by preventing gene flow between populations due to the presence of genetic incompatibilities, such as different RM systems, vector specificity, or negative epistasis. Other factors such as population dynamics and geographic range have been found to have an impact on speciation, although their relative contribution remains to be precisely deciphered. Overall, a BSC-based speciation model in prokaryotes would also rely on ecological processes and selection, as hypothesized by the SEM. However, one major difference with the SEM is that a BSC-based model of prokaryotic speciation predicts that speciation events can be driven by genetic incompatibilities and need not be systematically adaptive and ecologically-driven.

4 Species Borders and Pangenome Borders

Pangenome and Species Definitions The definition of species has direct consequences regarding the definition of pangenomes. If bacterial species are defined based on inconsistent criteria, it is not possible to compare the size of the pangenome across species and lineages. The case of *Prochlorococcus* illustrates this issue particularly well. *Prochlorococcus* is often studied as a single entity since it constitutes a single species based on 16S rRNA thresholds but multiple species based on ANI thresholds. The pangenome of *Prochlorococcus* has been estimated to reach the impressive amount of ~75,000 genes (Kashtan et al. 2014), although this would include strains that present less than 70% ANI, and this entity would actually correspond to multiple species and even genera. This issue likely affects many pangenome analyses considering that public databases frequently contain misclassified species and species classified based on inconsistent methods (Martiny et al. 2006; Comas et al. 2009; Trost et al. 2010). Studies focusing on the evolution of bacterial pangenomes should be based on rigorous species delimitation, since the misclassification of a single genome can lead to dramatic overestimates or underestimates of the size of a species' pangenome.

Species delimitation is not the only concern when analyzing pangenomes. The number of genomes sampled for each species obviously impacts pangenome estimates, since pangenomes necessarily increase in size as more genomes are included. It is possible to test if pangenome size reaches a plateau by performing resampling analyses, which would indicate that a sufficient number of genomes have been sampled to estimate the true pangenome size of the analyzed species (Tettelin et al. 2005; Lapierre and Gogarten 2009). Alternatively, it is possible to apply resampling analyses or to correct these metrics to account for uneven sampling biases across species (Bobay and Ochman 2018b). Biases in species sampling are a common issue for many genomic analyses and several methods have been developed as an attempt to address this shortcoming (Lapierre et al. 2016). However, the most efficient solution remains to increase sample sizes, and, more importantly, to limit biases when collecting samples, but this last consideration is often in conflict with study designs focusing on medically- or environmentally-relevant strains.

Cohesion of Core- and Pangenomes The goal of a species definition is to identify cohesive ensembles of evolutionary lineages. The ideal species definition would succeed in identifying genetically and ecologically cohesive units. Although genetic cohesion is easier to assess than ecological cohesion for bacteria, the genetic homogeneity of a group of organisms can be evaluated through different lenses. Firstly, because the core-genome constitutes the backbone of genes shared by all members of the species, these genes are more readily used to infer evolutionary relatedness and other metrics. Moreover, despite gene flow, core-genomes have conserved the phylogenetic signal of the vertical inheritance of bacterial taxa (Touchon et al. 2009; Abby et al. 2012). Nearly all genome-based species definitions—i.e., ANI, phylogenetic methods, and BSC-like—rely exclusively on the cohesion of the core-genome. The pangenome potentially offers an alternative measure of the genetic cohesion of species, since conspecific strains are expected to share more similar gene repertoires than strains belonging to distinct species. It is currently difficult to assess the pangenome cohesion of a species considering that accessory genes tend to be found at low frequency within species and this would require deep genome sampling, although more and more bacterial species have now hundreds or thousands of sequenced genomes. More analyses need to be performed to understand the specificity of pangenomes, especially in relation to closely related lineages and ecologically or geographically overlapping species.

Gene flow can define biological species based on DNA exchange along the core-genome but, so far, this method has been ignoring the patterns of HGT of the pangenome. The core- and pangenomes are two complementary metrics that can be used to infer the cohesion of species and some recent results obtained in two bacterial phyla suggest that core- and pangenomes present the same phylogenetic signal, implying that both can be reliable for inferring species borders (Wright and Baum 2018). In fact, a recent method has proposed a first attempt to delimitate species based on pangenome cohesion (Moldovan and Gelfand 2018), which opens promising possibilities to include pangenome cohesion into species delimitation. More work needs to be done in order to finely understand the evolutionary dynamics of the pangenome itself. For instance, the dynamics of the pangenome is likely affected by the ability of a given species to engage in gene flow, as suggested by a study showing that clonal species are unlikely to present a large pangenome, since their pangenome primarily evolves through gene loss (Bolotin and Hershberg 2015). Bacterial species can also gain new genes from external lineages and the extent of segregation of the pangenome remains poorly understood. The accumulation of genomic data should soon allow more accurate analysis of the dynamics of the pangenome and this will open new avenues for evaluating the genetic cohesion of prokaryotic species.

5 Drift-Barrier Model for Pangenome Evolution

A BSC-based species definition is particularly relevant for studying population genetics in prokaryotic organisms. Several parameters such as recombination rate, effective population size (*Ne*), or pangenome size are metrics that are typically inferred at the species level. In particular, *Ne* has strong implications regarding the relative impact of selection and drift acting on a given species. High *Ne* populations are less sensitive to drift and can efficiently purge deleterious sequences, whereas low *Ne* populations, on the other hand, will not be as effective at purging deleterious mutations. A trait conferred by a given variant would primarily evolve through drift (i.e., neutrally) when $|2.Ne.s| << 1$, while selection will be effective when $|2.Ne.s| >> 1$, where *s* represents the selection coefficient of a given sequence or variant (Kimura 1968). For these reasons, it is believed that more complex organisms such as mammals, which have low *Ne*, present larger genomes due to the accumulation of "junk DNA" through drift (i.e., the Mutational Hazard Hypothesis) (Lynch and Conery 2003; Lynch et al. 2011). Because these organisms display small population sizes, selection is not as efficient at purging slightly deleterious sequences, such as noncoding DNA, introns, and mobile elements.

In contrast to many eukaryotes, bacterial genomes are small and compact and because microbes present much larger population sizes, this seems in perfect agreement with the expectation of the Mutational Hazard hypothesis. The genomic compactness of bacteria has been ascribed to a strong bias toward deletion in these organisms (Mira et al. 2001; Andersson and Andersson 2001). However, several studies have observed that, across bacteria, genome size appears positively correlated with *Ne* (Daubin and Moran 2004; Kuo et al. 2009; Novichkov et al. 2009). Free-living bacteria frequently possess relatively large genomes (typically >3 Mb), while obligate endosymbionts—with low *Ne*—have smaller genomes (frequently

<1 Mb) (Moran and Plague 2004). Yet, some marine bacteria, which are thought to reach gigantic population sizes, also present streamlined genomes (Giovannoni et al. 2005, 2014). In particular, *Prochlorococcus* and *Pelagibacter ubique* have small genomes (~1 Mb), although they might be among the most abundant cellular organisms on earth (Batut et al. 2014). Therefore, the relationship between *Ne* and genome size appears to be more complex in bacteria.

One key difference between bacteria and higher eukaryotes is the very low amount of noncoding DNA, introns and mobile elements found in most bacterial genomes. In prokaryotes, variations in genome size are primarily driven by the presence of different amounts of accessory genes. Accessory genes are assumed to be functional and beneficial to the cell and recent modelling work suggests that virtually all genes in prokaryotic genomes are expected to be beneficial (Sela et al. 2016). Because the diversity of accessory genes is a direct function of pangenome size, this opens the possibility that *Ne* may drive the evolution of pangenome size rather than average genome size in prokaryotes. In support to this hypothesis, clear correlations between *Ne* and pangenome size have been observed across a dataset of 153 species, whose borders have been defined based on the BSC under a unified framework (Bobay and Ochman 2018b). Other recent studies have also reported similar trends (Mcinerney et al. 2017; Andreani et al. 2017).

Based on these observations, we have recently proposed that bacterial pangenomes could be driven by Drift-Barrier evolution (Bobay and Ochman 2018b). The Drift-Barrier model has originally been developed to account for the variation in mutation rates across organisms (Sung et al. 2012; Lynch et al. 2016). Under a Drift-Barrier model, pangenome size is expected to be a function of *Ne* because only the most beneficial accessory genes would be conserved by selection in small *Ne* species, while species with large *Ne* would be able to conserve accessory

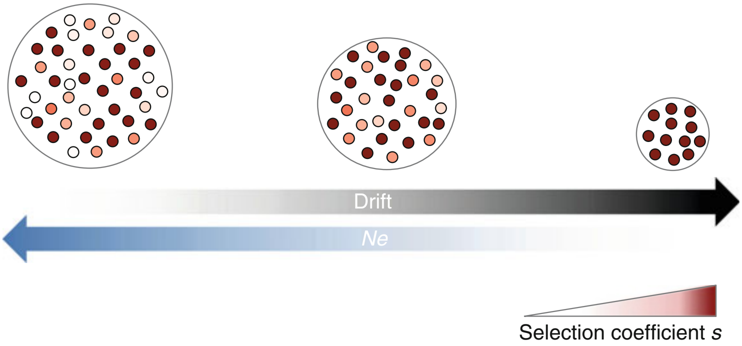


Fig. 1 Drift-Barrier model of pangenome evolution. Each large circle represents a pangenome and small circles represent individual genes. Color gradient reflects the selective coefficient of the genes. Species with large effective population size *Ne* are less subject to drift and can retain genes of small beneficial value (left). As *Ne* decreases, additional genes of small fitness benefit will be perceived as effectively neutral and will be lost by drift (center). Under strong levels of drift, as expected in small *Ne* species, only the most beneficial genes will be conserved by selection, and this will result in small pangenomes mostly composed of core/housekeeping genes (right)

genes with modest fitness contribution (Fig. 1). As supported by multiple studies, deleterious and neutral sequences are expected to be quickly purged from microbial genomes (Mira et al. 2001; Andersson and Andersson 2001). Our model assumes that virtually every gene of the pangenome is beneficial (positive selection coefficient: $s > 0$). Even if beneficial, an accessory gene is expected to be retained by selection only if it is perceived as *effectively* beneficial. In other words, an accessory gene will be conserved when $2.Ne.s >> 1$, while genes that appear effectively neutral ($2.Ne.s << 1$) are expected to be lost by drift. This implies that high *Ne* species are expected to retain a larger pool of genes including many accessory genes with modest fitness contribution, whereas low *Ne* species can only conserve the most beneficial genes (high *s*), i.e., mostly essential and/or core genes. Although new genes can be introduced into a species' pangenome by HGT, those accessory genes with low selective coefficient will be lost by drift.

6 Outlook

Many aspects of bacterial biology are now better understood but building a biologically-relevant microbial species concept remains challenging. Because prokaryotic organisms are microscopic, their population dynamics, ecological interactions, and speciation mechanisms are still difficult to decipher. Many aspects of the population processes driving microbial evolution have not been characterized. Habitat structure—and its temporal variations—of prokaryotic species is still for the large part mysterious. Similarly, microbial ecology and its impact on population dynamics remain tedious to describe in depth. Defining clear microbial niches is problematic practically and conceptually and little is known about microbial ecology compared to the vast collection of genomic data now available. The recent development of reverse ecology approaches opens a new route to gain knowledge about microbial ecology.

The accumulation of genomic data has profoundly impacted our vision of speciation in prokaryotic organisms. Several results suggest that prokaryotic species are definable and diagnosable as genetically cohesive as evidenced by the existence of a core-genome. However, the evolution of the core-genome remains to be fully understood. It is becoming possible to analyze the evolution of species- and genus-specific core-genomes over relatively short evolutionary time scales by comparing related species when sufficient genomic data is available (Touchon et al. 2014). On the other hand, the vast diversity of microbial pangenomes emphasizes the versatility of bacterial species. Much larger data sets are needed to accurately understand the dynamics of bacterial pangenomes, but several species now have thousands of sequenced genomes available. Deciphering the evolution of the pangenome will be highly insightful for our understanding of the dynamics and the genomic cohesion of microbial species.

From the original view of bacteria as purely clonal organisms, more and more evidence indicate that gene flow and HGT are key players in the evolution of most bacteria, and potentially act as major contributors to bacterial speciation. Computational approaches are needed to finely characterize gene flow in order to understand how networks of DNA routes can drive genomic cohesion and division in microbial species. Integrating these different aspects of bacterial biology will contribute to a more comprehensive prokaryotic species concept.

Acknowledgments I thank Bryan McLean, Eduardo PC Rocha, and Kasie Raymann for their opinion on the manuscript. LMB was supported by the National Science Foundation award DEB-1831730.

References

Abby SS, Tannier E, Gouy M, Daubin V (2012) Lateral gene transfer as a support for the tree of life. Proc Natl Acad Sci U S A 109:4962–4967

Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol 186:2629–2635

Andersson JO, Andersson SG (2001) Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. Mol Biol Evol 18:829–839

Andreani NA, Hesse E, Vos M (2017) Prokaryote genome fluidity is dependent on effective population size. ISME J 11(7):1719

Avery OT, MacLeod MC, McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of pneumonococcal types. Induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med 79:137–157

Bao YJ, Shapiro BJ, Lee SW, Ploplis VA, Castellino FJ (2016) Phenotypic differentiation of *Streptococcus pyogenes* populations is induced by recombination-driven gene-specific sweeps. Sci Rep 6:36644

Bapteste E, O'malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe FJ, Dupre J, Dagan T, Boucher Y, Martin W (2009) Prokaryotic evolution and the tree of life are two different things. Biol Direct 4:34

Batut B, Knibbe C, Marais G, Daubin V (2014) Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol 12:841–850

Bendall ML, Stevens SL, Chan LK, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG, Bertilsson S, Moran MA, Shade A, Newton RJ, Mcmahon KD, Malmstrom RR (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J 10:1589–1601

Bobay LM, Ochman H (2017a) Biological species are universal across life's domains. Genome Biol Evol 9:491–501

Bobay LM, Ochman H (2017b) Impact of recombination on the base composition of bacteria and archaea. Mol Biol Evol 34:2627–2636

Bobay LM, Ochman H (2018a) Biological species in the viral world. Proc Natl Acad Sci U S A 115:6040–6045

Bobay LM, Ochman H (2018b) Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol 18:153

Bobay LM, Rocha EP, Touchon M (2013) The adaptation of temperate bacteriophages to their host genomes. Mol Biol Evol 30:737–751

Bobay LM, Traverse CC, Ochman H (2015) Impermanence of bacterial clones. Proc Natl Acad Sci U S A 112:8893–8900

Bobay LM, Ellis BS, Ochman H (2018) ConSpeciFix: classifying prokaryotic species based on gene flow. Bioinformatics 21:3738–3740

Bolotin E, Hershberg R (2015) Gene loss dominates as a source of genetic variation within clonal pathogenic bacterial species. Genome Biol Evol 7:2173–2187

Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr CL, Polz MF (2011) Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. MBio 2:e00335

Brenner DJ, Staley J, Krieg N (2000) Classification of prokaryotic organisms and the concept of bacterial speciation. In: Boone DR, Castenholz RW, Garrity GM (eds) Bergey's manual of systematic biology, 2nd edn. Springer, New York

Brochier C, Forterre P, Gribaldo S (2005) An emerging phylogenetic core of archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. BMC Evol Biol 5:36

Buchrieser C, Glaser P, Rusniok C, Nedjari H, D'hauteville H, Kunst F, Sansonetti P, Parsot C (2000) The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of *Shigella flexneri*. Mol Microbiol 38:760–771

Cadillot-Quiroz H, Didelot X, Held N, Herrera A, Darling A, Reno M, Krause DJ, Whitaker RJ (2012) Patterns of gene flow define species of Thermophilic Archaea. PLoS Biol 10:e1001265

Caro-Quintero A, Konstantinidis KT (2012) Bacterial species may exist, metagenomics reveal. Environ Microbiol 14:347–355

Caro-Quintero A, Rodriguez-Castano GP, Konstantinidis KT (2009) Genomic insights into the convergence and pathogenicity factors of *Campylobacter jejuni* and *Campylobacter coli* species. J Bacteriol 191:5824–5831

Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287

Cohan FM (2001) Bacterial species and speciation. Syst Biol 50:513–524

Coleman ML, Chisholm SW (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A 107:18634–18639

Comas I, Homolka S, Niemann S, Gagneux S (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS One 4:e7815

Cordero OX, Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol 12:263–273

Cordero OX, Ventouras LA, Delong EF, Polz MF (2012) Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. Proc Natl Acad Sci U S A 109:20059–20064

Coyne JA, Orr HA (2004) Speciation. Sinauer Associates, Sunderland, MA

Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, Van Der Linden M, Mcgee L, Von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD (2011) Rapid pneumococcal evolution in response to clinical interventions. Science 331:430–434

Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD (2012) A high-resolution view of genome-wide pneumococcal transformation. PLoS Pathog 8:e1002745

Danchin EG, Rosso MN (2012) Lateral gene transfers have polished animal genomes: lessons from nematodes. Front Cell Infect Microbiol 2:27

Danner DB, Smith HO, Narang SA (1982) Construction of DNA recognition sites active in *Haemophilus* transformation. Proc Natl Acad Sci U S A 79:2393–2397

Daubin V, Moran NA (2004) Comment on "the origins of genome complexity". Science 306:978; author reply 978

David S, Sanchez-Buso L, Harris SR, Marttinen P, Rusniok C, Buchrieser C, Harrison TG, Parkhill J (2017) Dynamics and impact of homologous recombination on the evolution of *Legionella pneumophila*. PLoS Genet 13:e1006855

De Queiroz K (2007) Species concepts and species delimitation. Syst Biol 56:879–886

De Queiroz K, Gauthier J (1994) Toward a phylogenetic system of biological nomenclature. Trends Ecol Evol 9:27–31

Denef VJ, Mueller RS, Banfield JF (2010) AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. ISME J 4:599–610

Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–1266

Didelot X, Wilson DJ (2015) ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11:e1004041

Dobzhansky T (1935) A critique of the species concept in biology. Philos Sci 2:344–355

Doolittle WF, Papke RT (2006) Genomics and the bacterial species problem. Genome Biol 7:116

Doolittle WF, Zhaxybayeva O (2009) On the origin of prokaryotic species. Genome Res 19:744–756

Dorer MS, Sessler TH, Salama NR (2011) Recombination and DNA repair in *Helicobacter pylori*. Annu Rev Microbiol 65:329–348

Dykhuizen DE, Green L (1991) Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol 173:7257–7268

Escobar-Paramo P, Giudicelli C, Parsot C, Denamur E (2003) The evolutionary history of Shigella and enteroinvasive *Escherichia coli* revised. J Mol Evol 57:140–148

Everitt RG, Didelot X, Batty EM, Miller RR, Knox K, Young BC, Bowden R, Auton A, Votintseva A, Larner-Svensson H, Charlesworth J, Golubchik T, Ip CL, Godwin H, Fung R, Peto TE, Walker AS, Crook DW, Wilson DJ (2014) Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. Nat Commun 5:3956

Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. Science 315:476–480

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The bacterial species challenge: making sense of genetic and ecological diversity. Science 323:741–746

Friedman J, Alm EJ, Shapiro BJ (2013) Sympatric speciation: when is it possible in bacteria? PLoS One 8:e53539

Garrity GM (2016) A new genomics-driven taxonomy of bacteria and archaea: are we there yet? J Clin Microbiol 54:1956–1963

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309(5738):1242–1245

Giovannoni SJ, Cameron Thrash J, Temperton B (2014) Implications of streamlining theory for microbial ecology. ISME J 8:1553–1565

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19:2226–2238

Goodman SD, Scocca JJ (1988) Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. Proc Natl Acad Sci U S A 85:6982–6986

Griffith F (1928) The significance of pneumococcal types. J Hyg (Lond) 27:113–159

Groisman EA, Ochman H (1996) Pathogenicity islands: bacterial evolution in quantum leaps. Cell 87:791–794

Guglielmini J, De La Cruz F, Rocha EP (2013) Evolution of conjugation and type IV secretion systems. Mol Biol Evol 30:315–331

Hale TL, Keusch GT (1996) Shigella. In: Baron S (ed) Medical microbiology. University of Texas Medical Branch, Galveston, TX

Hanage WP (2013) Fuzzy species revisited. BMC Biol 11:41

Hanage WP (2016) Not so simple after all: bacteria, their population genetics, and recombination. Cold Spring Harb Perspect Biol 8(7):a018069

Hanage WP, Fraser C, Spratt BG (2005) Fuzzy species among recombinogenic bacteria. BMC Biol 3:6

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 8:11–22

Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner BL, Mori H, Horiuchi T (2006) Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. Mol Syst Biol 2:2006.0007

Hugenholtz P, Skarshewski A, Parks DH (2016) Genome-based microbial taxonomy coming of age. Cold Spring Harb Perspect Biol 8(6):a018085

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96:3801–3806

Johnston C, Martin B, Fichant G, Polard P, Claverys JP (2014) Bacterial transformation: distribution, shared mechanisms and divergent control. Nat Rev Microbiol 12:181–196

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science 344:416–420

Keller I, Wagner CE, Greuter L, Mwaiko S, Selz OM, Sivasundar A, Wittwer S, Seehausen O (2013) Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. Mol Ecol 22:2848–2863

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet 3:e231

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102:2567–2572

Konstantinidis KT, Rossello-Mora R, Amann R (2017) Uncultivated microbes in need of their own taxonomy. ISME J 11:2399–2406

Krause DJ, Whitaker RJ (2015) Inferring speciation processes from patterns of natural variation in microbial genomes. Syst Biol 64:926–935

Kuo CH, Moran NA, Ochman H (2009) The consequences of genetic drift for bacterial genome complexity. Genome Res 19:1450–1454

Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. Nat Rev Microbiol 8:317–327

Lang AS, Beatty JT (2007) Importance of widespread gene transfer agent genes in alphaproteobacteria. Trends Microbiol 15:54–62

Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. Trends Genet 25:107–110

Lapierre M, Blin C, Lambert A, Achaz G, Rocha EP (2016) The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. Mol Biol Evol 33:1711–1725

Lieberman TD, Flett KB, Yelin I, Martin TR, Mcadam AJ, Priebe GP, Kishony R (2014) Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. Nat Genet 46:82–87

Ludwig W, Klenk HP (2000) Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone DR, Castenholz RW, Garrity GM (eds) Bergey's manual of systematic biology, 2nd edn. Springer, New York

Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT (2011) Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci U S A 108:7200–7205

Lynch M, Conery JS (2003) The origins of genome complexity. Science 302:1401–1404

Lynch M, Bobay LM, Catania F, Gout JF, Rho M (2011) The repatterning of eukaryotic genomes by random genetic drift. Annu Rev Genomics Hum Genet 12:347–366

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL (2016) Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet 17:704–714

Maddamsetti R, Lenski RE (2018) Analysis of bacterial genomes from an evolution experiment with horizontal gene transfer shows that recombination can sometimes overwhelm selection. PLoS Genet 14:e1007199

Maddamsetti R, Lenski RE, Barrick JE (2015) Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. Genetics 200:619–631

Majewski J (2001) Sexual isolation in bacteria. FEMS Microbiol Lett 199:161–169

Majewski J, Cohan FM (1998) The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. Genetics 148:13–18

Majewski J, Cohan FM (1999) DNA sequence similarity requirements for interspecific recombination in *Bacillus*. Genetics 153:1525–1533

Majewski J, Zawadzki P, Pickerill P, Cohan FM, Dowson CG (2000) Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. J Bacteriol 182:1016–1023

Mallet J (2007) Hybrid speciation. Nature 446:279–283

Mallet J (2008) Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Philos Trans R Soc Lond Ser B Biol Sci 363:2971–2986

Mallet J, Beltran M, Neukirchen W, Linares M (2007) Natural hybridization in heliconiine butterflies: the species boundary as a continuum. BMC Evol Biol 7:28

Mallet J, Besansky N, Hahn MW (2016) How reticulated are species? BioEssays 38:140–149

Martiny JB, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Ovreas L, Reysenbach AL, Smith VH, Staley JT (2006) Microbial biogeography: putting microorganisms on the map. Nat Rev Microbiol 4:102–112

Marttinen P, Hanage WP (2017) Speciation trajectories in recombining bacterial species. PLoS Comput Biol 13:e1005640

Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J (2012) Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res 40:e6

Matic I, Taddei F, Radman M (2000) No genetic barriers between *Salmonella enterica* serovar typhimurium and *Escherichia coli* in SOS-induced mismatch repair-deficient cells. J Bacteriol 182:5922–5924

Mayr E (1942) Systematics and the origin of species. Columbia University Press, New-York

Mcinerney JO, Mcnally A, O'connell MJ (2017) Why prokaryotes have pangenomes. Nat Microbiol 2:17040

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15:589–594

Mell JC, Shumilina S, Hall IM, Redfield RJ (2011) Transformation of natural genetic variation into *Haemophilus influenzae* genomes. PLoS Pathog 7:e1002151

Meyer JR, Dobias DT, Medina SJ, Servilio L, Gupta A, Lenski RE (2016) Ecological speciation of bacteriophage lambda in allopatry and sympatry. Science 354:1301–1304

Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589–596

Moldovan MA, Gelfand MS (2018) Pangenomic definition of prokaryotic species and the phylogenetic structure of *Prochlorococcus* spp. Front Microbiol 9:428

Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A 93:2873–2878

Moran NA, Plague GR (2004) Genomic changes following host restriction in bacteria. Curr Opin Genet Dev 14:627–633

Moran NA, Mclaughlin HJ, Sorek R (2009) The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science 323:379–382

Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P (2017) Efficient inference of recent and ancestral recombination within bacterial populations. Mol Biol Evol 34:1167–1182

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS (2016) An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res 26:1612–1625

Novichkov PS, Wolf YI, Dubchak I, Koonin EV (2009) Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol 191:65–73

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Oliveira PH, Touchon M, Rocha EP (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res 42:10618–10631

Oliveira PH, Touchon M, Rocha EP (2016) Regulation of genetic flux between bacteria by restriction-modification systems. Proc Natl Acad Sci U S A 113:5658–5663

Papke RT, Zhaxybayeva O, Feil EJ, Sommerfeld K, Muise D, Doolittle WF (2007) Searching for species in haloarchaea. Proc Natl Acad Sci U S A 104:14092–14097

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36(10):996–1004

Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB, Huttenhower C, Morgan M, Segata N, Waldron L (2017) Accessible, curated metagenomic data through ExperimentHub. Nat Methods 14:1023–1024

Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet 29:170–175

Popa O, Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. Curr Opin Microbiol 14:615–623

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res 21:599–609

Popa O, Landan G, Dagan T (2017) Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. ISME J 11:543–554

Porter SS, Chang PL, Conow CA, Dunham JP, Friesen ML (2017) Association mapping reveals novel serpentine adaptation gene clusters in a population of symbiotic Mesorhizobium. ISME J 11:248–262

Pupo GM, Lan R, Reeves PR (2000) Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A 97:10567–10572

Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ (2009) Biogeography of the *Sulfolobus* islandicus pan-genome. Proc Natl Acad Sci U S A 106:8605–8610

Retchless AC, Lawrence JG (2007) Temporal fragmentation of speciation in bacteria. Science 317:1093–1096

Retchless AC, Lawrence JG (2010) Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proc Natl Acad Sci U S A 107:11453–11458

Richter M, Rossello-Mora R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106:19126–19131

Rieseberg LH (1997) Hybrid origins of plant species. Annu Rev Ecol Syst 28:359–389

Riley MA, Lizotte-Waniewski M (2009) Population genomics and the bacterial species concept. Methods Mol Biol 532:367–377

Roberts MS, Cohan FM (1993) The effect of DNA sequence divergence on sexual isolation in *Bacillus*. Genetics 134:401–408

Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, Mira A (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7:828–836

Rolland K, Lambert-Zechovsky N, Picard B, Denamur E (1998) Shigella and enteroinvasive *Escherichia coli* strains are derived from distinct ancestral strains of *E. coli*. Microbiology 144(Pt 9):2667–2672

Schubert S, Darlu P, Clermont O, Wieser A, Magistro G, Hoffmann C, Weinert K, Tenaillon O, Matic I, Denamur E (2009) Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species. PLoS Pathog 5:e1000257

Scocca JJ, Poland RL, Zoon KC (1974) Specificity in deoxyribonucleic acid uptake by transformable *Haemophilus influenzae*. J Bacteriol 118:369–373

Seehausen O (2004) Hybridization and adaptive radiation. Trends Ecol Evol 19:198–207

Sela I, Wolf YI, Koonin EV (2016) Theory of prokaryotic genome evolution. Proc Natl Acad Sci U S A 113:11399–11407

Shapiro BJ (2014) Signatures of natural selection and ecological differentiation in microbial genomes. Adv Exp Med Biol 781:339–359

Shapiro BJ, Polz MF (2014) Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol 22:235–247

Shapiro BJ, Polz MF (2015) Microbial speciation. Cold Spring Harb Perspect Biol 7:a018143

Shapiro BJ, David LA, Friedman J, Alm EJ (2009) Looking for Darwin's footprints in the microbial world. Trends Microbiol 17:196–204

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ (2012) Population genomics of early events in the ecological differentiation of bacteria. Science 336:48–51

Shapiro BJ, Leducq JB, Mallet J (2016) What is speciation? PLoS Genet 12:e1005860

Shen P, Huang HV (1986) Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics 112:441–457

Sheppard SK, Mccarthy ND, Falush D, Maiden MC (2008) Convergence of *Campylobacter* species: implications for bacterial evolution. Science 320:237–239

Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, Kelly DJ, Cody A, Colles FM, Strachan NJ, Ogden ID, Forbes K, French NP, Carter P, Miller WG, Mccarthy ND, Owen R, Litrup E, Egholm M, Affourtit JP, Bentley SD, Parkhill J, Maiden MC, Falush D (2013) Progressive genome-wide introgression in agricultural *Campylobacter coli*. Mol Ecol 22:1051–1064

Simmons SL, Dibartolo G, Denef VJ, Goltsman DS, Thelen MP, Banfield JF (2008) Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. PLoS Biol 6:e177

Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EP, De La Cruz F (2010) Mobility of plasmids. Microbiol Mol Biol Rev 74:434–452

Smith JM, Smith NH, O'rourke M, Spratt BG (1993) How clonal are bacteria? Proc Natl Acad Sci U S A 90:4384–4388

Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318:1449–1452

Stackebrandt E, Goebel BM (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Evol Microbiol 44:846–849

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M (2012) Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A 109:18488–18492

Syvanen M (2012) Evolutionary implications of horizontal gene transfer. Annu Rev Genet 46:341–358

Takeuchi N, Cordero OX, Koonin EV, Kaneko K (2015) Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. BMC Biol 13:20

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit Y, Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM (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

Thiergart T, Landan G, Martin WF (2014) Concatenated alignments and the case of the disappearing tree. BMC Evol Biol 14:266

Thingstad T, Lignell R (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquat Microb Ecol 13:19–27

Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3:711–721

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, El Karoui M, Frapy E, Garry L, Ghigo J, Gilles A, Johnson J, Le BouguéNec C, Lescat M, Mangenot S, Martinez-JéHanne V, Matic I, Nassif X, Oztas S, Petit M, Pichon C, Rouy Z, Saint Ruf C, Schneider D, Tourret J, Vacherie B, Vallenet D, MéDigue C, Rocha E, Denamur E (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 5:e1000344

Touchon M, Cury J, Yoon EJ, Krizova L, Cerqueira GC, Murphy C, Feldgarden M, Wortman J, Clermont D, Lambert T, Grillot-Courvalin C, Nemec A, Courvalin P, Rocha EP (2014) The genomic diversification of the whole *Acinetobacter* genus: origins, mechanisms, and consequences. Genome Biol Evol 6:2866–2882

Treangen TJ, Rocha EP (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet 7:e1001284

Trost B, Haakensen M, Pittet V, Ziola B, Kusalik A (2010) Analysis and comparison of the pan-genomic properties of sixteen well-characterized bacterial genera. BMC Microbiol 10:258

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N (2017) Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res 27:626–638

Velasco JD (2009) When monophyly is not enough: exclusivity as the key to defining a phylogenetic species concept. Biol Philos 24:473–486

Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154

Vos M (2011) A species concept for bacteria based on adaptive divergence. Trends Microbiol 19:1–7

Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. ISME J 3:199–208

Vulic M, Dionisio F, Taddei F, Radman M (1997) Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci U S A 94:9763–9767

Whitaker RJ, Grogan DW, Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science 301:976–978

Wiedenbeck J, Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35:957–976

Woese CR, Fox GE (1977) Phylogenetic structure of prokaryotic domain - primary kingdoms. Proc Natl Acad Sci U S A 74:5088–5090

Wright ES, Baum DA (2018) Exclusivity offers a sound yet practical species criterion for bacteria despite abundant gene flow. BMC Genomics 19:724

Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D (2014) Efficient inference of recombination hot regions in bacterial genomes. Mol Biol Evol 31:1593–1605

Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MC, Sheppard SK, Falush D (2015) The landscape of realized homologous recombination in pathogenic bacteria. Mol Biol Evol 33 (2):456–471

Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, Chun J (2017) Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol 67:1613–1617

Zawadzki P, Roberts MS, Cohan FM (1995) The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. Genetics 140:917–932

Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP (2009) Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. Genome Biol Evol 1:325–339

The Bacterial Guide to Designing a Diversified Gene Portfolio

Katherine A. Innamorati, Joshua P. Earl, Surya D. Aggarwal, Garth D. Ehrlich, and N. Luisa Hiller

Abstract The stunning ability of bacteria to evolve and adapt has contributed to the success of these single cells, which have inhabited the Earth for billions of years and play vital roles in the environment and in human health. The goal of this chapter is to present and discuss the population-level organizational scheme of bacterial pangenomes, wherein genes are distributed among the strains of a species, such that each individual strain encodes only a subset of the genes available at the population level. Genes from the accessory/distributed genome (those present only in a subset of strains within a species) impart diverse functions or variations on a conserved function to strains. Moreover, horizontal gene transfer generates novel gene combinations. The maintenance and spread of any given gene arrangement are influenced by fitness. Further, the extent of genomic plasticity is regulated by restriction modification systems, phage-defense systems, and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)—associated proteins (CRISPR-Cas). The combination of a pangenome structure and genomic plasticity reveals a successful strategy for bacterial adaptation to ever-changing environments. From a clinical perspective, pangenome analyses inform the selection of therapeutic targets, designed to focus either on an entire species or on virulence features within a species.

K. A. Innamorati · J. P. Earl
Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, PA, USA

Center for Genomic Sciences, Drexel University College of Medicine, Philadelphia, PA, USA

S. D. Aggarwal · N. L. Hiller (✉)
Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: lhiller@andrew.cmu.edu

G. D. Ehrlich (✉)
Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, PA, USA

Department of Otolaryngology - Head and Neck Surgery, Drexel University College of Medicine, Philadelphia, PA, USA

Center for Genomic Sciences, Drexel University College of Medicine, Philadelphia, PA, USA
e-mail: GE33@drexel.edu

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_3

Further, they provide a framework for modeling the efficacy of drugs and vaccines. In summary, following the explosion in sequencing technology, pangenome studies have revealed remarkable genomic organizations at the levels of species, with important implications to our understanding of evolution, and our ability to design therapeutics and predict their long-term outcomes.

Keywords Pangenome · Genomic diversity · Genomic plasticity · Horizontal gene transfer

1 Introduction

Bacteria dominate our planet and can be traced back to billions of years in the geological record. They play critical roles in shaping our habitat, from adding oxygen to the atmosphere to fixing nitrogen in the soil. They also play a vital role in human health, with commensal/mutualistic bacteria influencing nutrition and immunity, and pathogenic bacteria causing diseases from epidemics like the Black Death of medieval times to modern-day chronic biofilm infections resulting in the spread of antibiotic resistance. A defining characteristic of bacteria in both the environment and health is their ability to rapidly evolve and adapt. Here we discuss the elegant population-level organizational scheme that bacterial species use wherein their genomes are distributed among large numbers of strains, with no single strain having more than a small minority of genes available at the population level. This distributed pan(supra)-genome provides for adaptation to countless novel challenges and environmental niches.

Individual bacterial genomes have a discrete number of genes. However, enormous differences in gene content exist even among the genomes of strains of a single species. Therefore, the gene content of a single strain is less than the full complement of different genes from all strains. The comprehensive set of genes within a species, i.e., all genes from all strains, is defined as the ***pangenome*** (or supragenome). The pangenome is organized into the ***core genome***, which corresponds to the set of genes conserved across all strains in the species, and the ***accessory genome*** (or distributed genome), which are all noncore genes. We compiled pangenome papers from PubMed, identifying 295 species-specific pangenome projects performed on approximately 70 genera (Fig. 1). In all of these projects, the pangenome was found to be substantially larger than the core genome (Fig. 2).

The diversity within a species' pangenome provides a reservoir of genetic material available to bacterial cells to respond to selective pressures. Horizontal gene transfer (HGT) is the process by which individual bacterial cells can uptake genetic material from their environment or neighboring bacteria and generate novel, strain-specific gene combinations. It seems logical that when HGT occurs among strains of the same species these events are more likely to be adaptive or work in concert within the biological network, when compared to random mutations or genes acquired from distantly related species. This has been demonstrated to be the case in

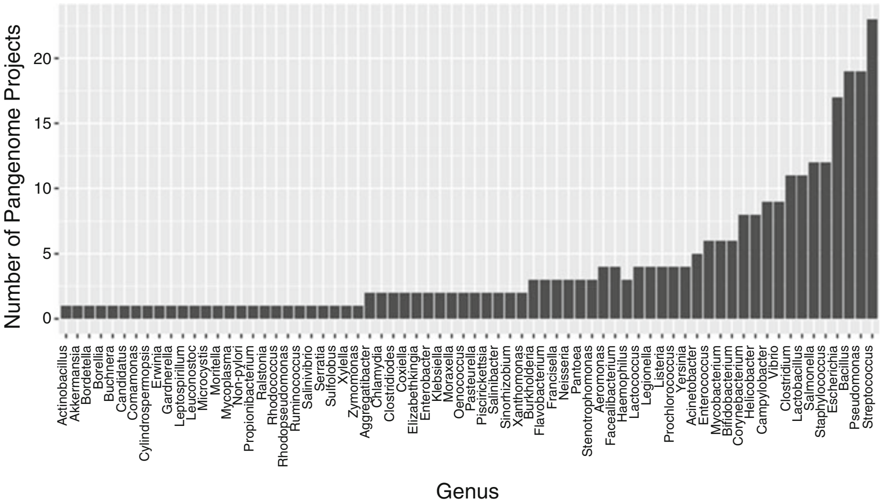


Fig. 1 Overview of the number of pangenomic studies per genus from a literature search between 2005 and 2018 (See "References for Fig. 1")

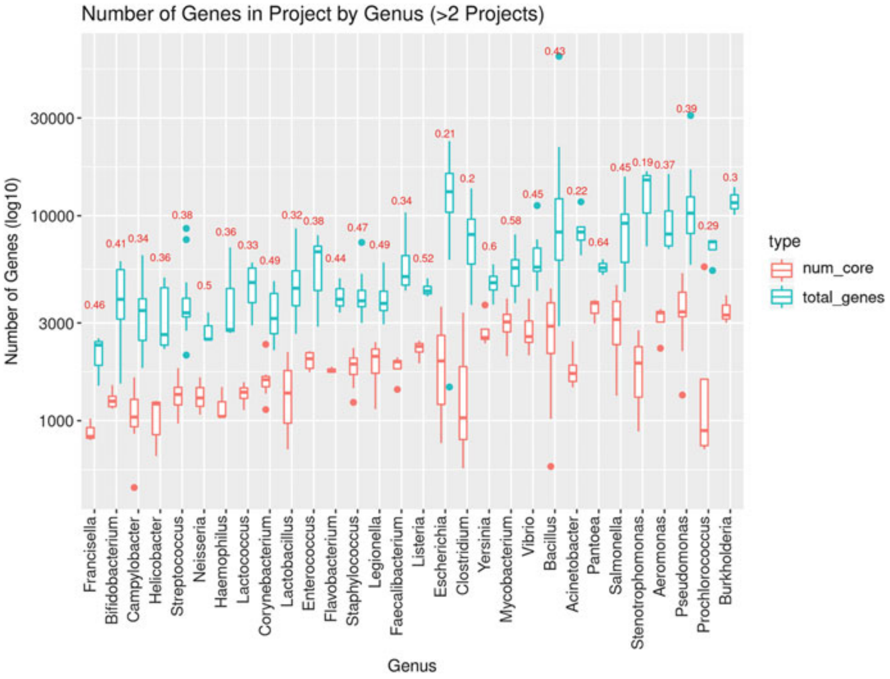


Fig. 2 The number of core and total gene clusters for genera with at least three available pangenomic projects. Numbers at the top correspond to the mean percent core. This is only an estimate as these numbers vary considerably based on parameters and strain selection processes

multiple species, where the majority of accessory genes appear to be evolving in tandem with the core genome (Gladitz et al. 2005). In this manner, the pangenome allows a species to incorporate more solutions to environmental stresses and niches than can be encoded by a single strain (Ehrlich et al. 2005, 2010).

2 Steps in the Assembly of a Pangenome

Pangenome analyses are performed on a set of strains from the same species, or very closely related species (often different species grouped together by genus, though we will not be examining those projects here). The set of all coding sequences (CDS) are clustered by sequence similarity with the objective of generating groups of orthologous genes. This is a multistep process that begins with whole-genome sequencing (WGS) of multiple independent bacterial (nonclonal, nonderivative) strains selected to represent the broadest geographic and phenotypic ranges of the species of interest. Following sequencing, the remaining steps are computational and include (1) assembly of genomes into contigs, (2) annotation of protein-coding sequences (CDS), and (3) clustering of CDSs based on the sequence similarity of nucleic acids or amino acids of their cognate encoded proteins. Once clusters are defined, they are classified based on strain prevalence into core or accessory (distributed) clusters. The accessory/distributed set of gene clusters is often further organized into those that are widely distributed (near core/soft core) in a population and those that are rare (shell) or unique (Fig. 3).

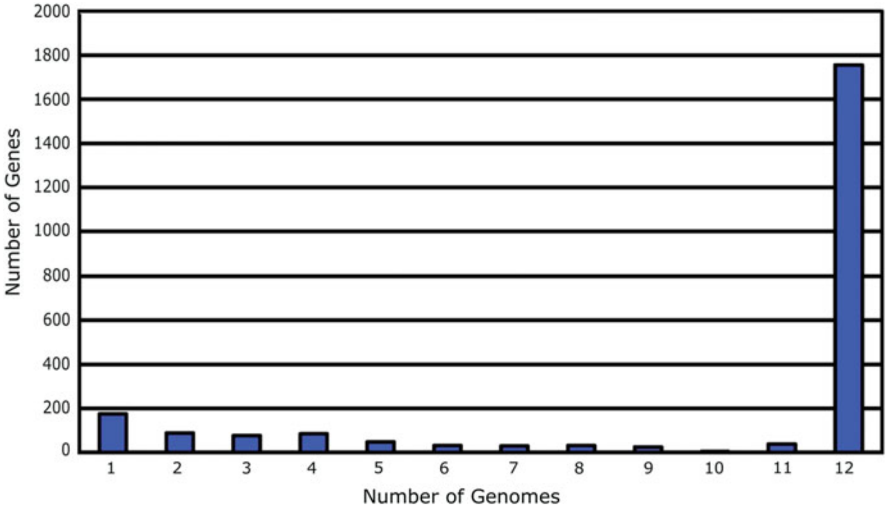


Fig. 3 Histogram of the number of gene clusters present in a given number of genomes. Taken from a project examining 12 genomes of *Moraxella catarrhalis* (Davie et al. 2011), with a total of 2383 gene clusters

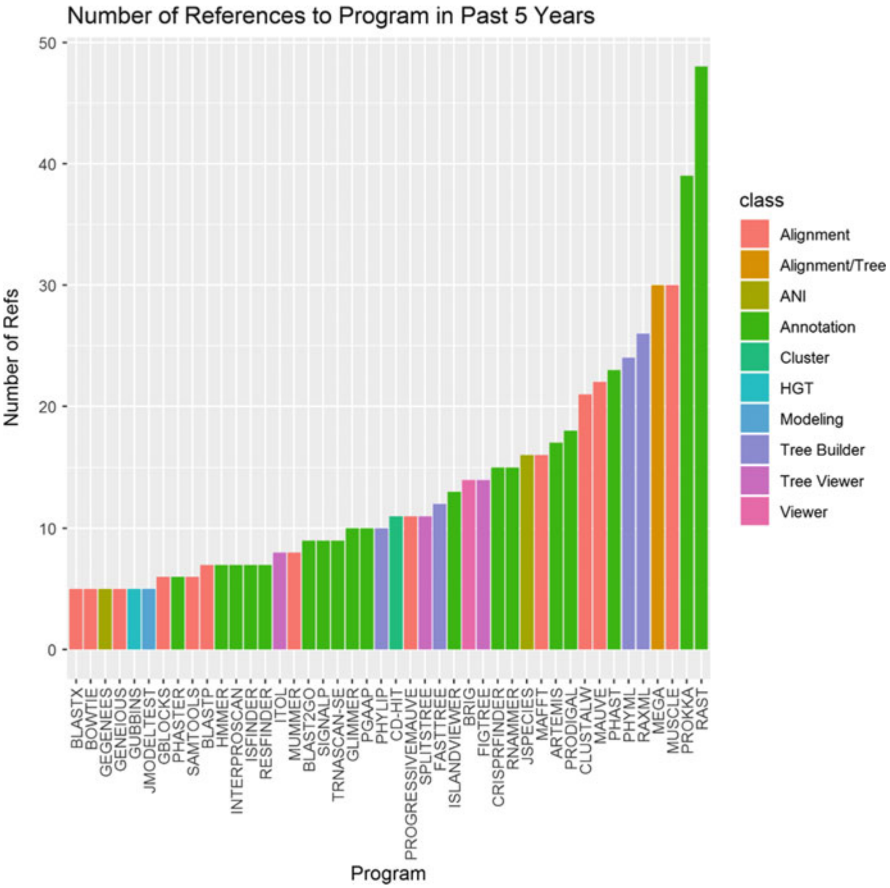


Fig. 4 Frequency of reference to programs over the past 5 years in pangenome publications (referenced at least 4 times)

The tools and the parameters used to characterize gene clusters vary widely among projects (Fig. 4). Generally, the first project(s) within a species tend to focus on the basic characterization of the pangenome. Subsequent projects often emphasize specific areas of interest, such as the distribution of virulence factors, levels of horizontal gene transfer, or epigenetic factors. Our survey of 295 pangenome projects did not reveal a strong preference for any individual assembly program. This is likely because assembly programs and versions perform differently depending on the examined species and the employed DNA sequencing technology. Further, many pangenome projects utilize pre-assembled genomes from publicly available databases (GenBank, EMBL, DDJB, JGI, PubMLST, etc.). This survey found that the CD-HIT program was the most frequently used gene clustering software, though a diverse set of other programs were also utilized for this purpose. Finally, commonly used software for other analyses include gene annotation (RAST, Prokka, PHAST, and Prodigal) (Aziz et al. 2008; Seemann 2014; Zhou et al. 2011; Hyatt et al. 2010), genome/gene alignments (Muscle, Mauve, Mega, and ClustalW) (Edgar 2004; Darling et al. 2004; Kumar et al. 1994; Higgins and Sharp 1988), and phylogenetic tree building (Mega, RAxML, and PhyML) (Kumar et al. 1994; Stamatakis 2006; Guindon et al. 2010). Overall, there is high variability in the methods/software used for pangenome analyses, reflecting diversity in the scope and goals of these projects.

3 Size of the Pangenome

The size of a species' pangenome, relative to the size of the core genome, is highly variable across the eubacteria. In Fig. 2, we display the variability we encountered in 295 species-specific pangenome projects (Figs. 1 and 2). Papers included in this summary span from 2005 [when the first pangenomes were described in *S. agalactiae* (Tettelin et al. 2005) and *H. influenzae* (Shen et al. 2005; Hogg et al. 2007)] through 2018. In all cases, the pangenome was significantly larger than the set of genes in a given strain. The size of the core genomes ranged from <20 to >60% of the pangenome (Fig. 2).

In some cases, calculations on the size of the pangenome may reflect inaccuracies in the current taxonomy, instead of the underlying biology. An instance of high genomic diversity is observed with *Gardnerella vaginalis*, where only 27% (746/2792) of its gene clusters are core (Ahmed et al. 2012). It is likely that *G. vaginalis* appears so genomically diverse because traditional biochemical tests used to identify strains within this taxa were unable to distinguish among the multiple genomically diverse species that are actually present. Thus, in this case, the apparent large size of the pangenome (and the corresponding small size of the core genome) arose from the unintentional merging of multiple species into a single species. In contrast, instances of low genomic diversity are observed in the genus *Bacillus*. Both *Bacillus anthracis* and *Bacillus thuringiensis* closely resemble *B. cereus* (Vilas-Bôas et al. 2007). *B. thuringiensis* appears to correspond to multiple phylogenetic clades (lineages) within *B. cereus*. *B. anthracis* (a species with one of the smallest pangenomes) likely represents a single phylogenetic lineage within the broader, more diverse definition of *B. cereus* that acquired a clinically important set of toxin genes (Okinaka and Keim 2016; Hall et al. 2010).

It is tempting to speculate that there are general principles that directly associate the size of the pangenome with the biology of the species. Factors that may play a substantial role are the extent of gene transfer, the degree of interactions with competing and cooperating species, the number of niches inhabited, or the lifestyle of the bacterium. The hypothesis that highly specialized environments lead to smaller genome sizes has been explored in the context of obligate intracellular species and pathogens (Merhej et al. 2009; Georgiades et al. 2011). A study of overall differences between the genomes of 12 highly pathogenic species compared to their most closely related nonpathogenic cousins found that, for the sets of bacteria studied, the most virulent species generally had smaller genomes, which suggests gene loss as well as loss-of-function mutations (Georgiades and Raoult 2011). The reduced genome size is hypothesized to be a consequence of extreme specialization of the pathogens to their hosts, while the less-specialized nonpathogens show greater levels of genomic variation due to selective pressure to remain competitive in more diverse environments (Georgiades and Raoult 2011). While this is an interesting idea, not all studies point to a relationship between pathogenicity and genome size (Bonar et al. 2018).

In a related vein, longitudinal comparative genomic studies of pathogenic clonal lineages of *Pseudomonas aeruginosa*, *Burkholderia* sp., and *Haemophilus influenzae* have captured microevolution and host adaptation in the human lung (Rau et al. 2012; Lee et al. 2017; Pettigrew et al. 2018; Moleres et al. 2018; Bianconi et al. 2018; Burns et al. 2001; Li et al. 2005; Jorth et al. 2015; Silva et al. 2016). In many cases, these changes reveal gene deletions when compared to their antecedents. For instance, serial isolates of *H. influenzae* clonal lineages in COPD patients display a significant association with loss-of-function mutations in the *ompP1* (*fadL*) accessory gene. *fadL* is beneficial to this bacterium in early infection, as it promotes adhesion and intracellular invasion via interactions with the epithelial cell ligand hCEACAM1 (human carcinoembryonic antigen-related cell adhesion molecule 1). In contrast, it may hinder long-term survival in the lung, as its expression increases sensitivity to arachidonic acid, an exogenous mammalian long-chain fatty acid with bactericidal effects (Moleres et al. 2018). This is indicative of selective pressure in favor of *ompP1* function in the nasopharynx and against its function in the lungs. These observations support the general concept that gene loss may accompany the ability to survive within highly circumscribed niches (Rau et al. 2012; Lee et al. 2017; Pettigrew et al. 2018; Moleres et al. 2018). Nonetheless, one must keep in mind that evolution in niches that do not support transmission may not be relevant to the evolution of the pangenome. Large-scale comparative pangenome and evolutionary studies promise to reveal the rules that shape the overall pangenome size, as well as identify disease and tissue-specific genes (and gene losses).

4 The Accessory Genome and Functional Diversity

In general, core genomes are enriched for housekeeping functions. These include energy production, amino acid metabolism, nucleotide metabolism, lipid transport, and translational machinery. Accessory genomes often encode genes involved in protein trafficking and defense, as well as many niche-specific functions. Further, plasmids, phage, and transposons are also often associated with accessory genomes. This section focuses on functional diversity as it pertains to the accessory genome.

Phenotypic traits can result from a blend of core genes with highly variable accessory genes. This is exemplified by the production of the capsule (Swartley et al. 1997; Bentley et al. 2006), synthesis of the extracellular polymeric substance

(EPS) (Harris et al. 2017), and modification of the cell wall (Gerlach et al. 2018). Here, conserved modules encoded in the core and softcore genomes are modified by components encoded by the accessory genome, providing a procedure to generate phenotypic variability. In *Neisseria meningitidis*, capsule biosynthesis genes are encoded within a single syntenic *cps* chromosomal region, which encodes both core and accessory genes. Variations in the accessory genes yield diversity in capsular types (Harrison et al. 2013). In *Lactobacillus salivarius*, the EPS cluster 2 contributes to the biofilm matrix. The genes at the extremities of this multigene cluster genes are core, while there is extensive variation in the genes encoded in the center of the cluster. These differences in glycotransferases and EPS biosynthesis-related proteins contribute to variations in the EPS structure (Harris et al. 2017). Yet another example is observed in methicillin-resistant *Staphylococcus aureus* (MRSA), where strains evade host immunity by modification of wall teichoic acid (WTA) using an alternative WTA glycosyltransferase encoded on a prophage (Gerlach et al. 2018). These studies exemplify how diversity within the accessory genome can provide bacteria with a blueprint to generate variability. This genomic flexibility is likely to increase the adaptive potential of bacterial species in the face of environmental stresses.

Genes encoded by the accessory genome can influence pathogenic potential. A well-studied example is *Escherichia coli*; this species encodes a highly diverse pangenome, where variability within the accessory genome leads to strains that differ in their ability to colonize human cell types and to trigger pathogenicity (Rasko et al. 2008). *E. coli* strains are grouped into pathovars based on the presence of virulence markers, often encoded on mobile elements (Kaper et al. 2004). Whole-genome comparative analyses of pathovars demonstrate that strains of the same pathovar are not always phylogenetically clustered (Rasko et al. 2008; Salipante et al. 2015; Hazen et al. 2013). This pattern of clustering is consistent with the transfer of accessory genes among *E. coli* strains, as well as the independent acquisition of virulence traits by strains in the same pathovar. One prominent example of HGT among *E. coli* strains of different pathovars is observed in the highly pathogenic strain that caused the 2011 German food poisoning outbreak (Mahan et al. 2013). Multiple genomic studies ultimately concluded that the outbreak was caused by a Shiga toxin-producing *E. coli* (STEC) of serotype O104:H4, which harbored multiple genes commonly associated with enteroaggregative *E. coli* (EAEC) including: a plasmid-encoded type I aggregative adherence fimbriae that mediate colonization and biofilm formation, assortment of serine proteases (SPATEs), and chromosomally encoded *Shigella* enterotoxin 1 (Askar et al. 2011; Mellmann et al. 2011; Rasko et al. 2011). Moreover, the prevalence of genetic transfer among *E. coli* strains is highlighted by the lack of an exclusive genomic signature among commensal *E. coli* strains. The strains that asymptomatically colonize the human gastrointestinal tract are genetically diverse (Rasko et al. 2008). These commensal strains may serve as genetic repositories for virulence determinants and, in addition, gene transfer events may modify their pathogenic potential and drug sensitivity. In conclusion, the accessory genome of *E. coli* is a critical determinant of tissue tropism, pathogenic potential, and clinical presentation.

Non-orthologous accessory genes with related functions are often syntenic across strains. We propose that this genomic configuration allows one variant to be switched by another in the process of recombination, where the neighboring genes provide an anchor for homologous recombination. One example is the genomic region that encodes the DpnI, DpnII, or the DpnIII type II restriction enzymes in *S. pneumoniae*. These loci differ in the sequence of the enzymes, the number of genes in the locus, and their ability to restrict phages or transforming DNA (Johnston et al. 2013a; Eutsey et al. 2015). Another example is the genomic region that encodes bacteriocins downstream of the *blp* histidine kinase signal transduction system in *S. pneumoniae*. While the genes in this region are predicted to be bacteriocins, the number of genes, their sequence, and the cells they target differ across strains (Lux et al. 2007; Dawid et al. 2007; Valente et al. 2016; Rezaei Javan et al. 2018). Other examples of this proposed mechanism, wherein conserved flanking genes anchor multiple variants of pathogenicity genes, include the parologous *vHiSLR* genes of *H. influenzae* (Kress-Bennett et al. 2016) and the *bro* gene variants of *Moraxella catarrhalis* (Earl et al. 2016). Syntenic regions that encode non-homologous genes within a single functional class may provide a pangenomic "switch," allowing cells to flip between variants of a single function to optimize fitness in diverse niches.

In summary, many of the genes in the accessory genome provide new functions or variations on a conserved function in a manner that expands the ability of strains to survive or adapt in their environments. In this manner, the strain diversity resulting from variations in the accessory genome may serve as a population-level tool to ensure the survival of a bacterial species.

5 Pangenome Plasticity

Speaking teleologically, via intra- and inter-species gene transfer, individual bacterial strains can draw from an expanded set of genes for their own adaptation and evolutionary success. This phenomenon was observed as early as 1928 in the Griffith's experiment, where a nonencapsulated strain of *S. pneumoniae* integrated DNA from an encapsulated isolate, leading to its conversion from avirulent to virulent (Griffith 1928). Almost a century later, the bacterial research community has described multitudinous instances of gene transfer among bacterial strains.

5.1 Gene Transfer Events Within and Across Species

Gene transfer events can occur anywhere, and our literature review identified 19 manuscripts that describe bacterial in vivo gene transfer within human patients (Table 1). A common theme is the acquisition of antibiotic resistance; particularly in regard to carbapenems, β-lactamases, and quinolones. Resistance was commonly the result of genes acquired via bacteriophages, plasmids, or pathogenicity islands

Table 1 Summary of studies on in vivo recombination

Bacterial species	Citation	Mechanism of transfer	Consequences and disease state
Acinetobacter baumannii	Agodi et al. (2006)	Class 1 integrons	ICU-acquired pneumonia multiresistant antibiotype
Enterobacteriaceae	Hammerum et al. (2016)	Plasmid	Meropenem resistance
Enterobacteriaceae	Datta et al. (2017)	Plasmid transfer of *bla*NDM-1	Septicemia
Enterobacter cloacae/Escherichia coli	Sidjabat et al. (2014)	Transfer of *bla*IMP-4	Meropenem resistance
Enterobacter aerogenes	Neuwirth et al. (2001)	Plasmid transfer-encoding ESBL TEM-24	Multidrug resistance
Escherichia coli	Soto et al. (2011)	Pathogenicity island acquisition	Male UTI recurrence
Escherichia coli	Schjørring et al. (2008), Bielaszewska et al. (2007)	Bacteriophage	Diarrhea and hemolytic uremic syndrome, gastroenteritis
Escherichia coli	Gumpert et al. (2017)	Conjugative antibiotic resistance plasmid	Antibiotic resistance
Haemophilus influenzae	Moleres et al. (2018)	Selective loss-of-function pressure	Loss of function-resistance to bactericidal fatty acids Acute COPD exacerbations
Klebsiella pneumoniae	Mena et al. (2006)	Insertion sequence (IS26)	Extended-spectrum beta-lactamase-producing species carbapenem resistance
Klebsiella pneumoniae/ Escherichia coli	Göttig et al. (2015)	Transconjugation of plasmid/transposon	Carbapenem resistance
Klebsiella pneumoniae/ Escherichia coli	Gona et al. (2014)	Mobile genetic elements carrying blaKPC, conjugative plasmids	Carbapenem-resistant patients developed bloodstream infections
Legionella pneumophila	McAdam et al. (2014)	Genomic island carrying T4SS	Legionnaires' disease/ community-acquired pneumonia-T4SS associated with more severe symptoms
Neisseria meningitidis	Brynildsrud et al. (2018)	Genomic islands, bacteriophage (MDAphi)	NmC meningitis
Serratia marcescens/ Escheric hia coli	Mata et al. (2010)	Plasmid mediated	AmpC beta-lactamase, quinolone resistance
Staphylococcus aureus/ epidermidis	Hurdle et al. (2005)	Conjugative replicon	Mupirocin resistance-persistent carrier of MRSA

(continued)

Table 1 (continued)

Bacterial species	Citation	Mechanism of transfer	Consequences and disease state
Staphylococcus aureus	Moore and Lindsay (2001)	Multiple mobile elements, specifically phages	Hospital MSSA
Staphylococcus aureus	Stanczak-Mrozek et al. (2015)	Bacteriophages and plasmids (general transduction)	Antibiotic-resistant MRSA
Staphylococcus aureus	Langhanki et al. (2018)	Mobile elements (genomic island, pathogenicity islands, bacteriophages), transduction	Long-term persistence cystic fibrosis patients

(Conlan et al. 2014; Bielaszewska et al. 2007; Datta et al. 2017; Feld et al. 2008; Langhanki et al. 2018; Mena et al. 2006; Neuwirth et al. 2001; Soto et al. 2011). In our set, five cases show HGT between different bacterial species: *Serratia marcescens* and *Escherichia coli* (Mata et al. 2010), two instances of *Klebsiella pneumoniae* and *E. coli* (Gona et al. 2014; Göttig et al. 2015), *Staphylococcus aureus* and *Staphylococcus epidermidis* (Hurdle et al. 2005), and *Enterobacter cloacae* and *E. coli* (Sidjabat et al. 2014). These studies highlight how bacteria occupying the same niche can evolve during the infectious disease process, posing new challenges for treatment.

Cross-species transfer events introduce new genes into the species, thus expanding the pangenome. A prominent example is acquisition of the type 3 secretion system (T3SS) by multiple Gram-negative bacteria. The T3SS allows for the transport of effector proteins from the bacterial cytosol directly into the host cells (Hacker et al. 1997; Hueck 1998). In most cases, the genes encoding this injection system, and their effectors, have been acquired by HGT (Brown and Finlay 2011). These T3SS systems are critical components of virulence. For instance, in *Salmonella*, acquisition of the SPI1 T3SS enables the bacterium to invade host cells, while acquisition of the SPI2 T3SS enables it to escape host defenses and survive within host cells inside a protective vacuole (Jennings et al. 2017; Ochman et al. 1996). Another example of cross-species transfer has been observed in *S. pneumoniae*, where a multigene locus was acquired from *Streptococcus suis* (Antic et al. 2017). This locus was acquired exclusively by a phylogenetically distinct subset of strains within the *S. pneumoniae* species—a subset much more likely to infect the conjunctiva. The genes acquired from *S. suis* appear to contribute to the tissue tropism by promoting adherence to the ocular epithelium. Thus, expansion of the pangenome by gene acquisition from outside the species can contribute to bacterial virulence and tropism.

Gene transfer among strains of the same species provides a mechanism to redistribute accessory/distributed genes within single strains. Studies on vaccine-escape strains of *S. pneumoniae* identified multiple genes acquired from a single donor (Golubchik et al. 2012). These recombination events ranged from 0.04 to

44 kb in size, and were located in various regions of the genome, including the capsular locus. Separate analyses of whole genomes of *S. pneumoniae* have captured multiple instances of serotype switches including from 23F to 3 and from 19F to 19A (Chewapreecha et al. 2014; Croucher et al. 2014a; Hiller et al. 2011). A current vaccine targets the 19F capsule, but not the 19A. Serotype 19F strains were widely prevalent pre-vaccine, while serotype 19A strains have spread in the USA during the post-vaccine era (Geno et al. 2015). This serotype switch has been observed in vaccinated and non-vaccinated populations. These observations are consistent with a model where HGT generates diverse genotypes, selective pressure from vaccines drives the spread of a subset of strains, and competition across strains shape the population and distribution of accessory genes.

Studies that describe recombination among strains driven by natural competence and transformation suggest that multiple transfers may occur both simultaneously and sequentially between individual donors and recipient strains. A study on *S. pneumoniae* captured the progressive accumulation of recombinations in a set of six clinical strains isolated from a pediatric patient over a 7-month period. One strain incurred multiple recombination events from the same donor, over two instances of recombination. These events introduced recombinations at 23 sites, and led to the exchange of over 7% of the genome (Hiller et al. 2010). Similarly, a laboratory study in *H. influenzae* also captured multiple gene transfer events after a bout of recombination (Mell et al. 2011). For this study, DNA from a clinical strain was used to transform a laboratory strain. Transformants were observed to have multiple recombination events over the length of the chromosome, collectively corresponding to ~1–3% of the genome. These analyses not only demonstrate HGT events across strains, but also suggest that strains may display multiple transfers during a single competence event.

HGT occurring through natural competence and transformation is unique among HGT mechanisms, in that it is driven by the recipient as opposed to by the donor (as is the case with mating and transduction). This means that it is an expressed phenotype that is triggered by the recipient cell. Thus, as a mechanism of mutation and evolution, it is expressed when a cell is stressed and provides a genetic means to adapt to a stressful environment resulting in mutation-on-demand (Ehrlich et al. 2005).

5.2 Constraints on Gene Transfer

While there is clear evidence of HGT among strains of the same species, distributed genes are not randomly distributed within a species. Instead, they tend to be associated with specific lineages, suggesting that pangenome evolution operates with forces that promote as well as limit gene transfer (Croucher et al. 2014b), as discussed in the next paragraphs.

There is increasing evidence that co-selection of genes limits gene transfer. A genome-wide study in *S. pneumoniae* demonstrated that a set of 876 loci, annotated to function in metabolism or transport, displayed a nonrandom distribution (Watkins et al. 2015). The authors show that groups of coevolved genes (alleles) are adapted to particular metabolic niches. They predict that disruption of these groups of alleles, a process mediated by HGT, would lead to a drop in strain fitness. A computational approach applied to *S. pneumoniae* and *N. meningitidis* also uncovered co-selection of genes associated with drug resistance and virulence (Pensar et al. 2019). Genome architecture may also limit gene transfer. Many bacterial genomes encode short sequences that are enriched in close proximity to the replication terminus. The location of these sequences is under selection, such that HGT events that disrupt these elements impose a fitness cost (Hendrickson et al. 2018). Thus, allele co-selection and genomic architecture illustrate genome-wide features that, when disturbed, can result in loss of fitness and consequently restrict gene flow.

In addition to factors that limit gene transfer via their influence on fitness, bacteria encode genes that serve as barriers to incoming DNA, such as restriction modification systems (RM), phage-defense systems, and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)—associated proteins (CRISPR-Cas). Most RM and CRISPR-Cas systems exert their influence on double-stranded DNA. While DNA entering the cell by transformation is single stranded, these systems still appear to serve as barriers to transformation; a compelling model proposed that they do so via their activity on the transformed chromosome (Johnston et al. 2013b). Studies in *N. meningitidis* and *S. pneumoniae* illustrate the role of restriction modification (RM) systems in limiting HGT. Strains of *N. meningitidis* organize into distinct phylogenetic groups that are associated with the distribution of >20 RM systems (Budroni et al. 2011). This distribution is consistent with the hypothesis that the RM systems limit HGT among clades. Similarly, the PMEN1 pandemic lineage of *S. pneumoniae* displays asymmetric gene transfer. The heterologous gene transfer from PMEN1 to other strains is abundant, yet into PMEN1 is modest (Wyres et al. 2012). The DpnIII RM system contributes to this structure, as it appears to limits HGT into PMEN1 strains, and is almost exclusively found in the PMEN1 lineage (Eutsey et al. 2015). Type I RM systems can also limit gene transfer, however, their architecture may allow rapid evolution of HGT barriers. The type I RM systems have a multifunctional component, where modification in one sequence can lead to both changes in methylation and endonuclease activity. This is in contrast to type II RM systems, where the protein that directs methylation is distinct from the protein that directs endonuclease activity, such that changes in specificity require mutations in more than one protein (Wilson and Murray 2003). In this manner, type I RM systems can rapidly evolve new specificities and generate diversity. A recent study in *S. pneumoniae* demonstrated that phase variation in the SpnIV phase-variable Type I RM limits acquisition of genomic islands by transformation (Kwun et al. 2018). The work captures an instance of phase variation on a type I RM system that generated an HGT barrier between nearly identical strains. Together, these studies suggest that RM systems may foster genomic stability within subsets of strains.

Many bacteria encode an abortive infection (Abi) system, which appears to be altruistic mechanism to protect the population at-large. When bacteria possessing an Abi system are infected by phage, the system is activated and triggers the death of the bacterial host. In this manner, death of the infected isolate avoids spread of the phage across the bacterial community (Chopin et al. 2005). In an exciting twist, phage defense systems may also be encoded by prophage, illustrating cooperation between bacteria and phage to restrict unrelated phages (Dedrick et al. 2017; Bondy-Denomy et al. 2016).

CRISPR-Cas confers adaptive immunity in prokaryotes and has the ability to inhibit conjugation, transduction and transformation. The CRISPR-Cas are composed of arrays of palindromic nucleotide repeats that are interspersed by short unique DNA segments called spacers, and *cas* genes. The spacers are acquired from foreign DNA, usually bacteriophages. Following acquisition, spacers are transcribed and processed into small CRISPR RNA (crRNA) molecules. A complex formed by Cas proteins and crRNA leads to the degradation of invading foreign nucleic acid, protecting cells from future invasion (Jiang and Doudna 2017; Adli 2018). Many bacterial species and lineages are devoid of CRISPR-Cas systems. In vitro studies in multiple bacteria reveal an inverse correlation between HGT and the presence of a functional CRISPR-Cas system (Jiang et al. 2013; Watson et al. 2018). In *Enterococcus faecalis*, multidrug-resistant plasmids were observed in strains that lacked CRISPR-Cas systems, while the drug-sensitive strains encoded this system (Palmer and Gilmore 2010). Further, under selective pressure for the acquisition of antibiotic-resistant plasmids, *Staphylococcus epidermidis* strains acquired inactivating mutations in the CRISPR-Cas system (Jiang et al. 2013). These studies suggest that bacteria encounter a tradeoff: the fitness advantages associated with phage resistance afforded by CRISPR-Cas must be balanced against a decrease in genomic plasticity and the benefits conferred by acquisition of novel genes. Nonetheless, the role of phage protection systems in restricting gene flow is far from fully resolved. Some studies find contrasting results, and do not support the conclusion that CRISPR-Cas limits HGT. A large-scale computational study revealed that the activity of the CRISPR-Cas system was not associated with HGT events over long evolutionary timescales (Gophna et al. 2015). Further, a study in *Pectobacterium atrosepticum* suggests that CRISPR-Cas systems may actually contribute to HGT via their role in protecting bacteria against phage attack (Watson et al. 2018). Thus, more research is required to determine the ultimate influence of CRISPR-Cas systems on the genomic plasticity of bacterial populations.

In conclusion, the set of genes in a species' pangenome can expand via the introduction of genes from other species, rearrange across strains via an intra-species exchange, or vary with mutations. The shuffling of accessory genes and alleles generates new combinations that are subsequently subjected to the forces of selection on gene products and genome-wide features. Moreover, RMs, CRISPR-Cas, and phage-defense systems may also influence gene flow across strains and species. All factors combined, genomic plasticity emerges as a successful strategy for bacterial survival.

6 A Balance in the Accessory Genome

A remarkable observation comes from recent mathematical models and population studies. Negative frequency-dependent selection may stabilize the proportion of individual accessory genes in a population of *S. pneumoniae* (Azarian et al. 2018; Corander et al. 2017). As expected, the authors observed that vaccination led to a dramatic drop in the representation of vaccine-sensitive strains. In doing so, the distribution of accessory genes within the population differed from that of the pre-vaccine population. Interestingly, over time, the frequency of the accessory genes trended toward that seen in the pre-vaccine population. These results suggest that the distribution of genes in the pneumococcal pangenome may have an equilibrium point. It remains to be determined whether similar patterns are observed in other species. The suggestion that the composition of pangenomes tends toward an equilibrium has important implications regarding our ability to predict the nature of replacement strains after the introduction of therapies that target subsets of strains within a bacterial population using a microbiome-sparing approach.

7 Clinical Applications

Pangenomic analyses can be utilized to identify potential therapeutic targets. Target specificity can be customized depending on the desired effect. The core genome can be used to target an entire species, as it contains genes possessed by every member of the species. Alternatively, targeting select members of the accessory genome, or the "microbiome-sparing" approach, will ensure that only strains containing the gene of interest are affected. Both strategies can be utilized to combat a wide variety of pathogens.

Current efforts to combat pathogenic bacteria include targeting the bacterial capsule, a large polysaccharide layer that is a major virulence determinant with a key role in immune evasion. Strains vary in the composition of their capsules: those with identical capsules are placed in the same serotype, and those with highly similar capsules within a serogroup. For example, there are over 97 different serotypes known for *S. pneumoniae* that fall into 46 serogroups (Bentley et al. 2006; Geno et al. 2015; Tzeng et al. 2016), and over 12 serotypes for *N. meningitidis* (Harrison et al. 2013; Geno et al. 2015; Tzeng et al. 2016; Claus et al. 1997). New serotypes can arise by HGT, like in the movement of *SiaD* genes between *N. meningitidis* strains, or through mispairing during gene replication, which is responsible for serotypes 15 B/C in *S. pneumoniae* (Claus et al. 1997; van Selm et al. 2003). Capsular polysaccharide vaccines are available for *S. pneumoniae*, *S. typhi*, and *N. meningitidis* (Geno et al. 2015; Tzeng et al. 2016; Hessel et al. 1999). These specifically target the bacterial capsule, but young children (under the age of two) fail to create antibodies against these vaccines. To combat this, polysaccharide–protein conjugate vaccines were designed, which combine the polysaccharide antigen with protein carriers and render them more immunogenic in young children (Finn 2004; Nair 2012; Szu et al. 1989; Lin et al. 2001). Development of conjugate vaccines faces major challenges, such as cost, host immune response, and bacterial structures (Nair 2012). Therefore, it would be ideal to create capsular polysaccharide vaccines with better immunogenicity. However, the structures of some capsule sugars are too similar to those found in mammalian tissues to be useful as polysaccharide vaccines. In these cases, vaccines could be designed to target virulence via accessory genes or to target these species as a whole via the core genome (Pichichero 2017; Daniels et al. 2016; Chan et al. 2018).

Using the accessory genome to create strain-specific drugs and vaccines has wide implications. For example, it is easy to imagine the creation of therapies against bacterial pathogens that are able to spare the larger microbiome. Commensal bacteria in the microbiome and pathogenic bacteria of the same species may share the same core genome, but can have vast differences in the content of their accessory genomes. If a therapy targets protein products from genes found only in the accessory genomes of pathogenic bacteria, it will not disturb the patient’s microflora as the commensal bacteria would lack the proteins the therapy is created against. This strategy has the potential to greatly improve patient health and recovery following a bacterial infection.

Pangenomic studies can aid in the development of diagnostic tools. As with vaccines and drug development, accessory genes can be used to identify a particular strain/phenotype and core genes to identify a specific species. A study of 17 clinical isolates of *G. vaginalis* was used to propose the reclassification of *G. vaginalis* as a genus, based on the extent of pangenomic variation (Ahmed et al. 2012). Previously, metronidazole was used as a blanket antibiotic for the treatment of bacterial vaginosis. However, the understanding that metronidazole-resistant clades of *G. vaginalis* are actually different species creates room for the development of diagnostic tools to inform antibiotic treatment for patients with bacterial vaginosis (Balashov et al. 2014). Similarly, pangenomic studies among phenotypically divergent *M. catarrhalis* strains led to the characterization of a deep phylogenetic clade structure that separated the pathogenic sero-resistant strains from commensal sero-sensitive strains (Earl et al. 2016). In yet another example, *Staphylococcus epidermidis* was divided into two phylogenetic groups. One group included both commensals and pathogens, the other composed exclusively of commensal strains. Strains in the second group-encoded formate dehydrogenase, revealing a potential diagnostic marker (Conlan et al. 2012). A study in *Helicobacter pylori* identified lineage-specific genes; some have already been associated with acid resistance and virulence, and thus are potential targets to guide treatments (van Vliet 2017). Moreover, when studies associating pangenome and phenotype identify unannotated genes as diagnostic markers, they provide genetic fodder for linking new functions, distribution, and disease outcome (Ehrlich et al. 2010). One caution to consider in the development of diagnostics is that chronic infections can be caused by multiple strains of the same species, and analysis of a single strain could misdirect treatment.

A crucial benefit of pangenomic analyses is their ability to determine the presence or absence of antibiotic-resistant markers. Prescription of an ineffective antibiotic is both detrimental to patient's health and adds to the problem of global antibiotic resistance. Some examples of pangenomic analyses to study the distribution and transmission of resistance genes have been performed on *E. coli* strains collected from wastewater treatment plants (Mahfouz et al. 2018), community-associated *Clostridium difficile* strains isolated from farm animals and humans (Knetsch et al. 2018), and strains of *Stenotrophomonas maltophilia* collected from cystic fibrosis (CF) patients (Esposito et al. 2017). Given that related strains often differ in their drug resistance profile, probing the accessory genome for genes that encode drug resistance will be a critical component of personalized medicine.

Genome-scale models (GEMs) of metabolism can provide great insight into the link between metabolism and pathogenesis. These network reconstructions provide context for the relationship between gene, gene product, and phenotype. Pangenomic analyses in three species observed that the majority of core genes are associated with metabolism (Cornejo et al. 2013; Bosi et al. 2016; Vieira et al. 2011). Pangenomic analysis of inflammatory bowel disease (IBD)-associated *E. coli* strains reported metabolic differences between IBD-associated strains and nonassociated strains, where the former set appeared to utilize energy more efficiently (Fang et al. 2018). The differences in metabolic capabilities in disease and healthy states provide a promising place to explore diagnostic applications of the pangenome. Furthermore, the link between metabolism and virulence can be explored, and be used diagnostically to differentiate strains that cause mild or severe symptom presentation (Bosi et al. 2016).

Beyond the use of pangenomic analyses to select targets for vaccines, therapeutics, and diagnosis, it has also served as an epidemiological tool. The origin of the 2010 cholera outbreak in Haiti was traced using pangenomic analysis of *Vibrio cholerae*. Initially, it was unclear whether the epidemic originated with a local strain or Asian strain. A pangenomic analysis revealed that the epidemic was caused by strains originated in Southeast Asia (Reimer et al. 2011; Hendriksen et al. 2011; Chin et al. 2011; Mutreja et al. 2011; Orata et al. 2014; Hasan et al. 2012). Such epidemiological studies allow better strategic planning to avoid future epidemics.

8 Conclusions

The Distributed Genome Hypothesis provides both a historical and theoretical framework for understanding bacterial genomic plasticity, and puts it in the context of other classes of chronic pathogens (viruses and eukaryotic parasites) that have developed different mechanistic strategies for the generation of genetic diversity in situ. Viruses such as HIV-1 utilize an error-prone DNA polymerase (reverse transcriptase) to generate enormous diversity resulting in the development of a quasispecies within days of infection (Korber et al. 2001). Trypanosomes utilize a cassetting mechanism for antigen switching wherein they have an entire chromosome of outer surface protein cassettes that they can exchange within the larger functional protein whenever the host adaptive immune response recognizes the previous cassette (Horn 2014). Thus, within this context, we can view HGT of distributed genes among bacterial strains of a species as yet another means of "programmed" variation (Ehrlich et al. 2010).

9 Perspectives

The plasticity provided by the eubacterial pangenome may be driving the evolution of other domains of life. The rapid recombination of bacterial strains provided the evolutionary pressure for the development of the vertebrate adaptive immune system—which is mechanistically similar to what the bacteria are doing—it is essentially a random gene rearrangement phenomenon, very similar to HGT (Hu et al. 2007). Lastly, as the variability in species becomes apparent, it triggers the question of how best to define a species. While pangenomic analyses do not offer the ultimate solution, they may provide a useful definition. Once the core genome of a species is defined, strains can be assigned, or not assigned, to a species based on the extent to which they share the same core genome (Nistico et al. 2014).

References

Adli M (2018) The CRISPR tool kit for genome editing and beyond. Nat Commun 9:1911

Agodi A, Zarrilli R, Barchitta M, Anzaldi A, Di Popolo A, Mattaliano A, Ghiraldi E, Travali S (2006) Alert surveillance of intensive care unit-acquired Acinetobacter infections in a Sicilian hospital. Clin Microbiol Infect 12(3):241–247

Ahmed A, Earl J, Retchless A, Hillier SL, Rabe LK, Cherpes TL et al (2012) Comparative genomic analyses of 17 clinical isolates of *Gardnerella vaginalis* provide evidence of multiple genetically isolated clades consistent with subspeciation into genovars. J Bacteriol 194:3922–3937

Antic I, Brothers KM, Stolzer M, Lai H, Powell E, Eutsey R et al (2017) Gene acquisition by a distinct phyletic group within *Streptococcus pneumoniae* promotes adhesion to the ocular epithelium. mSphere 2:e00213. https://doi.org/10.1128/mSphere.00213-17

Askar M, Faber MS, Frank C, Bernard H, Gilsdorf A, Fruth A, et al (2011) Update on the ongoing outbreak of haemolytic uraemic syndrome due to Shiga toxin-producing *Escherichia coli* (STEC) serotype O104, Germany, May 2011. Euro Surveill 16. Available https://www.ncbi.nlm.nih.gov/pubmed/21663710

Azarian T, Grant LR, Arnold BJ, Hammitt LL, Reid R, Santosham M et al (2018) The impact of serotype-specific vaccination on phylodynamic parameters of *Streptococcus pneumoniae* and the pneumococcal pan-genome. PLoS Pathog 14:e1006966

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA et al (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75

Balashov SV, Mordechai E, Adelson ME, Gygax SE (2014) Identification, quantification and subtyping of *Gardnerella vaginalis* in noncultured clinical vaginal samples by quantitative PCR. J Med Microbiol 63:162–175

Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M et al (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet 2:e31

Bianconi I, D'Arcangelo S, Esposito A, Benedet M, Piffer E, Dinnella G et al (2018) Persistence and microevolution of *Pseudomonas aeruginosa* in the cystic fibrosis lung: a single-patient longitudinal genomic study. Front Microbiol 9:3242

Bielaszewska M, Prager R, Köck R, Mellmann A, Zhang W, Tschäpe H et al (2007) Shiga toxin gene loss and transfer in vitro and in vivo during enterohemorrhagic *Escherichia coli* O26 infection in humans. Appl Environ Microbiol 73:3144–3150

Bonar EA, Bukowski M, Hydzik M, Jankowska U, Kedracka-Krok S, Groborz M et al (2018) Joint genomic and proteomic analysis identifies meta-trait characteristics of virulent and non-virulent strains. Front Cell Infect Microbiol 8:313

Bondy-Denomy J, Qian J, Westra ER, Buckling A, Guttman DS, Davidson AR et al (2016) Prophages mediate defense against phage infection through diverse mechanisms. ISME J 10:2854–2866

Bosi E, Monk JM, Aziz RK, Fondi M, Nizet V, Palsson BØ (2016) Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. Proc Natl Acad Sci U S A 113:E3801–E3809

Brown NF, Finlay BB (2011) Potential origins and horizontal transfer of type III secretion systems and effectors. Mob Genet Elem 1:118–121

Brynildsrud OB, Eldholm V, Bohlin J, Uadiale K, Obaro S, Caugant DA (2018) Acquisition of Virulence genes by a carrier strain gave rise to the ongoing epidemics of meningococcal disease in West Africa. Proc Natl Acad Sci U S A 115(21):5510–5515

Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C et al (2011) *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci U S A 108:4494–4499

Burns JL, Gibson RL, McNamara S, Yim D, Emerson J, Rosenfeld M et al (2001) Longitudinal assessment of *Pseudomonas aeruginosa* in young children with cystic fibrosis. J Infect Dis 183:444–452

Chan W-Y, Entwisle C, Ercoli G, Ramos-Sevillano E, McIlgorm A, Cecchini P et al (2018) A novel, multiple-antigen pneumococcal vaccine protects against lethal *Streptococcus pneumoniae* challenge. Infect Immun 87(3):e00846. https://doi.org/10.1128/IAI.00846-18

Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L et al (2014) Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet 46:305–309

Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR et al (2011) The origin of the Haitian cholera outbreak strain. N Engl J Med 364:33–42

Chopin M-C, Chopin A, Bidnenko E (2005) Phage abortive infection in Lactococci: variations on a theme. Curr Opin Microbiol 8:473–479

Claus H, Vogel U, Mühlenhoff M, Gerardy-Schahn R, Frosch M (1997) Molecular divergence of the sia locus in different serogroups of *Neisseria meningitidis* expressing polysialic acid capsules. Mol Gen Genet 257:28–34

Conlan S, Mijares LA, NISC Comparative Sequencing Program, Becker J, Blakesley RW, Bouffard GG et al (2012) *Staphylococcus epidermidis* pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. Genome Biol 13:R64

Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP et al (2014) Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. Sci Transl Med 6:254ra126

Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD et al (2017) Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nat Ecol Evol 1:1950–1960

Cornejo OE, Lefébure T, Bitar PDP, Lang P, Richards VP, Eilertson K et al (2013) Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. Mol Biol Evol 30:881–893

Croucher NJ, Chewapreecha C, Hanage WP, Harris SR, McGee L, van der Linden M et al (2014a) Evidence for soft selective sweeps in the evolution of pneumococcal multidrug resistance and vaccine escape. Genome Biol Evol 6:1589–1602

Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP (2014b) Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Commun 5:5471

Daniels CC, Rogers PD, Shelton CM (2016) A review of pneumococcal vaccines: current polysaccharide vaccine recommendations and future protein antigens. J Pediatr Pharmacol Ther 21:27–35

Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403

Datta S, Mitra S, Chattopadhyay P, Som T, Mukherjee S, Basu S (2017) Spread and exchange of bla NDM-1 in hospitalized neonates: role of mobilizable genetic elements. Eur J Clin Microbiol Infect Dis 36:255–265

Davie JJ, Earl J, de Vries SPW, Ahmed A, Hu FZ, Bootsma HJ et al (2011) Comparative analysis and supragenome modeling of twelve *Moraxella catarrhalis* clinical isolates. BMC Genomics 12:70

Dawid S, Roche AM, Weiser JN (2007) The blp bacteriocins of *Streptococcus pneumoniae* mediate intraspecies competition both in vitro and in vivo. Infect Immun 75:443–451

Dedrick RM, Jacobs-Sera D, Bustamante CAG, Garlena RA, Mavrich TN, Pope WH et al (2017) Prophage-mediated defence against viral attack and viral counter-defence. Nat Microbiol 2:16251

Earl JP, de Vries SPW, Ahmed A, Powell E, Schultz MP, Hermans PWM et al (2016) Comparative genomic analyses of the *Moraxella catarrhalis* serosensitive and seroresistant lineages demonstrate their independent evolution. Genome Biol Evol 8:955–974

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Ehrlich GD, Hu FZ, Shen K, Stoodley P, Post JC (2005) Bacterial plurality as a general mechanism driving persistence in chronic infections. Clin Orthop Relat Res 437:20–24

Ehrlich GD, Ahmed A, Earl J, Hiller NL, Costerton JW, Stoodley P et al (2010) The distributed genome hypothesis as a rubric for understanding evolution in situ during chronic bacterial biofilm infectious processes. FEMS Immunol Med Microbiol 59:269–279

Esposito A, Pompilio A, Bettua C, Crocetta V, Giacobazzi E, Fiscarelli E et al (2017) Evolution of *Stenotrophomonas maltophilia* in cystic fibrosis lung over chronic infection: a genomic and phenotypic population study. Front Microbiol 8:1590

Eutsey RA, Powell E, Dordel J, Salter SJ, Clark TA, Korlach J et al (2015) Genetic stabilization of the drug-resistant PMEN1 *Pneumococcus* lineage by its distinctive DpnIII restriction-modification system. MBio 6:e00173

Fang X, Monk JM, Mih N, Du B, Sastry AV, Kavvas E et al (2018) *Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. BMC Syst Biol 12:66

Feld L, Schjørring S, Hammer K, Licht TR, Danielsen M, Krogfelt K et al (2008) Selective pressure affects transfer and establishment of a *Lactobacillus plantarum* resistance plasmid in the gastrointestinal environment. J Antimicrob Chemother 61:845–852

Finn A (2004) Bacterial polysaccharide-protein conjugate vaccines. Br Med Bull 70:1–14

Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C et al (2015) Pneumococcal capsules and their types: past, present, and future. Clin Microbiol Rev 28:871–899

Georgiades K, Raoult D (2011) Genomes of the most dangerous epidemic bacteria have a virulence repertoire characterized by fewer genes but more toxin-antitoxin modules. PLoS One 6:e17962

Georgiades K, Merhej V, El Karkouri K, Raoult D, Pontarotti P (2011) Gene gain and loss events in *Rickettsia* and *Orientia* species. Biol Direct 6:6

Gerlach D, Guo Y, De Castro C, Kim S-H, Schlatterer K, Xu F-F et al (2018) Methicillin-resistant *Staphylococcus aureus* alters cell wall glycosylation to evade immunity. Nature 563:705–709

Gladitz J, Shen K, Antalis P, Hu FZ, Post JC, Ehrlich GD (2005) Codon usage comparison of novel genes in clinical isolates of *Haemophilus influenzae*. Nucleic Acids Res 33:3644–3658

Golubchik T, Brueggemann AB, Street T, Gertz RE, Spencer CCA, Ho T et al (2012) Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. Nat Genet 44:352–355

Gona F, Barbera F, Pasquariello AC, Grossi P, Gridelli B, Mezzatesta ML et al (2014) In vivo multiclonal transfer of bla(KPC-3) from *Klebsiella pneumoniae* to *Escherichia coli* in surgery patients. Clin Microbiol Infect 20:O633–O635

Gophna U, Kristensen DM, Wolf YI, Popa O, Drevet C, Koonin EV (2015) No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. ISME J 9:2021–2027

Göttig S, Gruber TM, Stecher B, Wichelhaus TA, Kempf VAJ (2015) In vivo horizontal gene transfer of the carbapenemase OXA-48 during a nosocomial outbreak. Clin Infect Dis 60:1808–1815

Griffith F (1928) The significance of *Pneumococcal* types. J Hyg 27:113–159

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321

Gumpert H, Kubicek-Sutherland JZ, Porse A, Karami N, Munck C, Linkevicius M, Adlerberth I, Wold AE, Andersson DI, Sommer MOA (2017) Transfer and persistence of a multi-drug resistance plasmid in situ of the infant gut microbiota in the absence of antibiotic treatment. Front Microbiol 8(September):1852

Hacker J, Blum-Oehler G, Mühldorfer I, Tschäpe H (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 23:1089–1097

Hall BG, Ehrlich GD, Hu FZ (2010) Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology 156:1060–1068

Hammerum AM, Hansen F, Nielsen HL, Jakobsen L, Stegger M, Andersen PS, Jensen P et al (2016) Use of WGS data for investigation of a long-term NDM-1-producing *Citrobacter freundii* outbreak and secondary in vivo spread of blaNDM-1 to *Escherichia coli*, *Klebsiella pneumoniae* and *Klebsiella oxytoca*. J Antimicrob Chemother 71(11):3117–3124

Harris HMB, Bourin MJB, Claesson MJ, O'Toole PW (2017) Phylogenomics and comparative genomics of, a mammalian gut commensal. Microb Genom 3:e000115

Harrison OB, Claus H, Jiang Y, Bennett JS, Bratcher HB, Jolley KA et al (2013) Description and nomenclature of *Neisseria meningitidis* capsule locus. Emerg Infect Dis 19:566–573

Hasan NA, Choi SY, Eppinger M, Clark PW, Chen A, Alam M et al (2012) Genomic diversity of 2010 Haitian cholera outbreak strains. Proc Natl Acad Sci U S A 109:E2010

Hazen TH, Sahl JW, Fraser CM, Donnenberg MS, Scheutz F, Rasko DA (2013) Refining the pathovar paradigm via phylogenomics of the attaching and effacing *Escherichia coli*. Proc Natl Acad Sci U S A 110:12810–12815

Hendrickson HL, Barbeau D, Ceschin R, Lawrence JG (2018) Chromosome architecture constrains horizontal gene transfer in bacteria. PLoS Genet 14:e1007421

Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM et al (2011) Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. MBio 2:e00157

Hessel L, Debois H, Fletcher M, Dumas R (1999) Experience with Salmonella typhi Vi capsular polysaccharide vaccine. Eur J Clin Microbiol Infect Dis 18:609–620

Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73:237–244

Hiller NL, Ahmed A, Powell E, Martin DP, Eutsey R, Earl J et al (2010) Generation of genic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PLoS Pathog 6:e1001108

Hiller NL, Eutsey RA, Powell E, Earl JP, Janto B, Martin DP et al (2011) Differences in genotype and virulence among four multidrug-resistant *Streptococcus pneumoniae* isolates belonging to the PMEN1 clone. PLoS One 6:e28850

Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R et al (2007) Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol 8:R103

Horn D (2014) Antigenic variation in African trypanosomes. Mol Biochem Parasitol 195:123–129

Hu FZ, Hogg J, Hiller NL, Janto B, Boissy R, Post JC, Ehrlich GD (2007) Biofilms as bacterial breeding grounds: a counterpoint to the adaptive host response. Abstract MS15 9th international symposium on recent advances in otitis media, June 3–7, 2007, in St. Pete Beach, FL

Hueck CJ (1998) Type III protein secretion systems in bacterial pathogens of animals and plants. Microbiol Mol Biol Rev 62:379–433

Hurdle JG, O'Neill AJ, Mody L, Chopra I, Bradley SF (2005) In vivo transfer of high-level mupirocin resistance from *Staphylococcus epidermidis* to methicillin-resistant *Staphylococcus aureus* associated with failure of mupirocin prophylaxis. J Antimicrob Chemother 56:1166–1168

Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf 11:119

Jennings E, Thurston TLM, Holden DW (2017) Salmonella SPI-2 type III secretion system effectors: molecular mechanisms and physiological consequences. Cell Host Microbe 22:217–231

Jiang F, Doudna JA (2017) CRISPR-Cas9 structures and mechanisms. Annu Rev Biophys 46:505–529

Jiang W, Maniv I, Arain F, Wang Y, Levin BR, Marraffini LA (2013) Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. PLoS Genet 9:e1003844

Johnston C, Polard P, Claverys J-P (2013a) The DpnI/DpnII pneumococcal system, defense against foreign attack without compromising genetic exchange. Mob Genet Elem 3:e25582

Johnston C, Martin B, Polard P, Claverys J-P (2013b) Postreplication targeting of transformants by bacterial immune systems? Trends Microbiol 21:516–521

Jorth P, Staudinger BJ, Wu X, Hisert KB, Hayden H, Garudathri J et al (2015) Regional isolation drives bacterial diversification within cystic fibrosis lungs. Cell Host Microbe 18:307–319

Kaper JB, Nataro JP, Mobley HL (2004) Pathogenic *Escherichia coli*. Nat Rev Microbiol 2:123–140

Knetsch CW, Kumar N, Forster SC, Connor TR, Browne HP, Harmanus C et al (2018) Zoonotic transfer of *Clostridium difficile* harboring antimicrobial resistance between farm animals and humans. J Clin Microbiol 56(3):e01384. https://doi.org/10.1128/JCM.01384-17

Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V (2001) Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull 58:19–42

Kress-Bennett JM, Hiller NL, Eutsey RA, Powell E, Longwell MJ, Hillman T et al (2016) Identification and characterization of msf, a novel virulence factor in *Haemophilus influenzae*. PLoS One 11:e0149891

Kumar S, Tamura K, Nei M (1994) MEGA: molecular evolutionary genetics analysis software for microcomputers. Comput Appl Biosci 10:189–191

Kwun MJ, Oggioni MR, De Ste Croix M, Bentley SD, Croucher NJ (2018) Excision-reintegration at a pneumococcal phase-variable restriction-modification locus drives within- and between-strain epigenetic differentiation and inhibits gene acquisition. Nucleic Acids Res 46:11438–11453

Langhanki L, Berger P, Treffon J, Catania F, Kahl BC, Mellmann A (2018) In vivo competition and horizontal gene transfer among distinct *Staphylococcus aureus* lineages as major drivers for adaptational changes during long-term persistence in humans. BMC Microbiol 18:152

Lee AH-Y, Flibotte S, Sinha S, Paiero A, Ehrlich RL, Balashov S et al (2017) Phenotypic diversity and genotypic flexibility of *Burkholderia cenocepacia* during long-term chronic infection of cystic fibrosis lungs. Genome Res 27:650–662

Li Z, Kosorok MR, Farrell PM, Laxova A, West SEH, Green CG et al (2005) Longitudinal development of mucoid *Pseudomonas aeruginosa* infection and lung disease progression in children with cystic fibrosis. JAMA 293:581–588

Lin FY, Ho VA, Khiem HB, Trach DD, Bay PV, Thanh TC et al (2001) The efficacy of a *Salmonella typhi* Vi conjugate vaccine in two-to-five-year-old children. N Engl J Med 344:1263–1269

Lux T, Nuhn M, Hakenbeck R, Reichmann P (2007) Diversity of bacteriocins and activity spectrum in *Streptococcus pneumoniae*. J Bacteriol 189:7741–7751

Mahan MJ, Kubicek-Sutherland JZ, Heithoff DM (2013) Rise of the microbes. Virulence 4:213–222

Mahfouz N, Caucci S, Achatz E, Semmler T, Guenther S, Berendonk TU et al (2018) High genomic diversity of multi-drug resistant wastewater *Escherichia coli*. Sci Rep 8:8928

Mata C, Miró E, Mirelis B, Garcillán-Barcia MP, de la Cruz F, Coll P et al (2010) In vivo transmission of a plasmid coharbouring bla and qnrB genes between *Escherichia coli* and *Serratia marcescens*. FEMS Microbiol Lett 308:24–28

McAdam PR, Vander Broek CW, Lindsay DSJ, Ward MJ, Hanson MF, Gillies M, Watson M, Stevens JM, Edwards GF, Fitzgerald JR (2014) Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. Genome Biol 15(11):504

Mell JC, Shumilina S, Hall IM, Redfield RJ (2011) Transformation of natural genetic variation into *Haemophilus influenzae* genomes. PLoS Pathog 7:e1002151

Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, 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

Mena A, Plasencia V, García L, Hidalgo O, Ayestarán JI, Alberti S et al (2006) Characterization of a large outbreak by CTX-M-1-producing *Klebsiella pneumoniae* and mechanisms leading to in vivo carbapenem resistance development. J Clin Microbiol 44:2831–2837

Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol Direct 4:13

Moleres J, Fernández-Calvet A, Ehrlich RL, Martí S, Pérez-Regidor L, Euba B et al (2018) Antagonistic pleiotropy in the bifunctional surface protein FadL (OmpP1) during adaptation of *Haemophilus influenzae* to chronic lung infection associated with chronic obstructive pulmonary disease. MBio 9:e01176. https://doi.org/10.1128/mBio.01176-18

Moore PC, Lindsay JA (2001) Genetic variation among hospital isolates of methicillin-sensitive *Staphylococcus aureus*: evidence for horizontal transfer of Virulence genes. J Clin Microbiol 39 (8):2760–2767

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S et al (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462–465

Nair M (2012) Protein conjugate polysaccharide vaccines: challenges in development and global implementation. Indian J Community Med 37:79–82

Neuwirth C, Siebor E, Pechinot A, Duez JM, Pruneaux M, Garel F et al (2001) Evidence of in vivo transfer of a plasmid encoding the extended-spectrum beta-lactamase TEM-24 and other resistance factors among different members of the family Enterobacteriaceae. J Clin Microbiol 39:1985–1988

Nistico L, Earl J, Hiller L, Ahmed A, Retchless A, Janto B, Costerton JC, Hu FZ, Ehrlich GD (2014) Using the core and supra genomes to determine diversity and natural proclivities among bacterial strains. In: Skovhus TL, Caffrey S, Hubert C (eds) Application of molecular microbiological methods. Caister Academic Press, Norfolk, UK, p 1

Ochman H, Soncini FC, Solomon F, Groisman EA (1996) Identification of a pathogenicity island required for Salmonella survival in host cells. Proc Natl Acad Sci 93:7800–7804

Okinaka RT, Keim P (2016) The phylogeny of *Bacillus cereus* sensu lato. Microbiol Spectr 4(1). https://doi.org/10.1128/microbiolspec.TBS-0012-2012

Orata FD, Keim PS, Boucher Y (2014) The 2010 cholera outbreak in Haiti: how science solved a controversy. PLoS Pathog 10:e1003967

Palmer KL, Gilmore MS (2010) Multidrug-resistant enterococci lack CRISPR-cas. MBio 1:e00227. https://doi.org/10.1128/mBio.00227-10

Pensar J, Puranen S, MacAlasdair N, Kuronen J, Tonkin-Hill G, Pesonen M et al (2019) Genome-wide epistasis and co-selection study using mutual information. Genomics. bioRxiv

Pettigrew MM, Ahearn CP, Gent JF, Kong Y, Gallo MC, Munro JB et al (2018) *Haemophilus influenzae* genome evolution during persistence in the human airways in chronic obstructive pulmonary disease. Proc Natl Acad Sci U S A 115:E3256–E3265

Pichichero ME (2017) Pneumococcal whole-cell and protein-based vaccines: changing the paradigm. Expert Rev Vaccines 16:1181–1190

Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P et al (2008) The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol 190:6881–6893

Rasko DA, Webster DR, Sahl JW, 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

Rau MH, Marvig RL, Ehrlich GD, Molin S, Jelsbak L (2012) Deletion and acquisition of genomic content during early stage adaptation of *Pseudomonas aeruginosa* to a human host environment. Environ Microbiol 14:2200–2211

Reimer AR, 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

Rezaei Javan R, van Tonder AJ, King JP, Harrold CL, Brueggemann AB (2018) Genome sequencing reveals a large and diverse repertoire of antimicrobial peptides. Front Microbiol 9:2012

Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM et al (2015) Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. Genome Res 25:119–128

Schjørring S, Struve C, Krogfelt KA (2008) Transfer of antimicrobial resistance plasmids from *Klebsiella pneumoniae* to *Escherichia coli* in the mouse intestine. J Antimicrob Chemother 62 (5):1086–1093

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069

Shen K, Antalis P, Gladitz J, Sayeed S, Ahmed A, Yu S et al (2005) Identification, distribution, and expression of novel genes in 10 clinical isolates of nontypeable *Haemophilus influenzae*. Infect Immun 73:3479–3491

Sidjabat HE, Heney C, George NM, Nimmo GR, Paterson DL (2014) Interspecies transfer of blaIMP-4 in a patient with prolonged colonization by IMP-4-producing Enterobacteriaceae. J Clin Microbiol 52:3816–3818

Silva IN, Santos PM, Santos MR, Zlosnik JEA, Speert DP, Buskirk SW et al (2016) Long-term evolution of *Burkholderia* multivorans during a chronic cystic fibrosis infection reveals shifting forces of selection. mSystems 1:e00029. https://doi.org/10.1128/mSystems.00029-16

Soto SM, Zúñiga S, Ulleryd P, Vila J (2011) Acquisition of a pathogenicity island in an *Escherichia coli* clinical isolate causing febrile urinary tract infection. Eur J Clin Microbiol Infect Dis 30:1543–1550

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690

Stanczak-Mrozek KI, Manne A, Knight GM, Gould K, Witney AA, Lindsay JA (2015) Within-host diversity of MRSA antimicrobial resistances. J Antimicrob Chemother 70(8):2191–2198

Swartley JS, Marfin AA, Edupuganti S, Liu LJ, Cieslak P, Perkins B et al (1997) Capsule switching of *Neisseria meningitidis*. Proc Natl Acad Sci U S A 94:271–276

Szu SC, Li XR, Schneerson R, Vickers JH, Bryla D, Robbins JB (1989) Comparative immunogenicities of Vi polysaccharide-protein conjugates composed of cholera toxin or its B subunit as a carrier bound to high- or lower-molecular-weight Vi. Infect Immun 57:3823–3827

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL 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

Tzeng Y-L, Thomas J, Stephens DS (2016) Regulation of capsule in *Neisseria meningitidis*. Crit Rev Microbiol 42:759–772

Valente C, Dawid S, Pinto FR, Hinds J, Simões AS, Gould KA et al (2016) The blp locus of *Streptococcus pneumoniae* plays a limited role in the selection of strains that can Cocolonize the human nasopharynx. Appl Environ Microbiol 82:5206–5215

van Selm S, van Cann LM, Kolkman MAB, van der Zeijst BAM, van Putten JPM (2003) Genetic basis for the structural difference between *Streptococcus pneumoniae* serotype 15B and 15C capsular polysaccharides. Infect Immun 71:6192–6198

van Vliet AHM (2017) Use of pan-genome analysis for the identification of lineage-specific genes of *Helicobacter pylori*. FEMS Microbiol Lett 364(2):fnw296. https://doi.org/10.1093/femsle/fnw296

Vieira G, Sabarly V, Bourguignon P-Y, Durot M, Le Fèvre F, Mornico D et al (2011) Core and panmetabolism in *Escherichia coli*. J Bacteriol 193:1461–1472

Vilas-Bôas GT, Peruca APS, Arantes OMN (2007) Biology and taxonomy of *Bacillus cereus*, *Bacillus anthracis*, and *Bacillus thuringiensis*. Can J Microbiol 53:673–687

Watkins ER, Penman BS, Lourenço J, Buckee CO, Maiden MCJ, Gupta S (2015) Vaccination drives changes in metabolic and virulence profiles of *Streptococcus pneumoniae*. PLoS Pathog 11:e1005034

Watson BNJ, Staals RHJ, Fineran PC (2018) CRISPR-Cas-mediated phage resistance enhances horizontal gene transfer by transduction. MBio 9:e02406. https://doi.org/10.1128/mBio.02406-17

Wilson GG, Murray NE (2003) Restriction and modification systems. Annual Reviews. Palo Alto, CA. https://doi.org/10.1146/annurev.ge.25.120191.003101

Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, Liñares J et al (2012) The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success. Genome Biol 13:R103

Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS (2011) PHAST: a fast phage search tool. Nucleic Acids Res 39:W347–W352

References for Fig. 1

Abreu VAC, Popin RV, Alvarenga DO, Schaker PDC, Hoff-Risseti C, Varani AM, Fiore MF (2018) Genomic and genotypic characterization of *Cylindrospermopsis raciborskii*: toward an intraspecific phylogenetic evaluation by comparative genomics. Front Microbiol 9 (February):306

Alhashash F, Wang X, Paszkiewicz K, Diggle M, Zong Z, McNally A (2016) Increase in bacteraemia cases in the east midlands region of the UK due to MDR *Escherichia coli* ST73: high levels of genomic and plasmid diversity in causative isolates. J Antimicrob Chemother 71 (2):339–343

Ali A, Soares SC, Santos AR, Guimarães LC, Barbosa E, Almeida SS, Abreu VAC et al (2012) Campylobacter fetus subspecies: comparative genomics and prediction of potential virulence targets. Gene 508(2):145–156

Ali A, Naz A, Soares SC, Bakhtiar M, Tiwari S, Hassan SS, Hanan F et al (2015) Pan-genome analysis of human gastric pathogen *H. pylori*: comparative genomics and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets. Biomed Res Int:139580

Anastasi E, MacArthur I, Scortti M, Alvarez S, Giguère S, Vázquez-Boland JA (2016) Pangenome and phylogenomic analysis of the pathogenic actinobacterium *Rhodococcus equi*. Genome Biol Evol 8(10):3140–3148

Andreani NA, Carraro L, Martino ME, Fondi M, Fasolato L, Miotto G, Magro M, Vianello F, Cardazzo B (2015) A genomic and transcriptomic approach to investigate the blue pigment phenotype in *Pseudomonas fluorescens*. Int J Food Microbiol 213(November):88–98

Arboleya S, Bottacini F, O'Connell-Motherway M, Ryan CA, Ross RP, van Sinderen D, Stanton C (2018) Gene-trait matching across the *Bifidobacterium longum* pan-genome reveals considerable diversity in carbohydrate catabolism among human infant strains. BMC Genomics 19(1):33

Argemi X, Matelska D, Ginalski K, Riegel P, Hansmann Y, Bloom J, Pestel-Caron M, Dahyot S, Lebeurre J, Prévost G (2018a) Comparative genomic analysis of *Staphylococcus lugdunensis* shows a closed pan-genome and multiple barriers to horizontal gene transfer. BMC Genomics 19(1):621

Argemi X, Nanoukon C, Affolabi D, Keller D, Hansmann Y, Riegel P, Baba-Moussa L, Prévost G (2018b) Comparative genomics and identification of an enterotoxin-bearing pathogenicity island, SEPI-1/SECI-1, in *Staphylococcus epidermidis* pathogenic strains. Toxins 10(3):93. https://doi.org/10.3390/toxins10030093

Asenjo F, Olmos A, Henríquez-Piskulich P, Polanco V, Aldea P, Ugalde JA, Trombert AN (2016) Genome sequencing and analysis of the first complete genome of *Lactobacillus kunkeei* strain MP2, an *Apis mellifera* gut isolate. PeerJ 4(April):e1950

Baddam R, Kumar N, Shaik S, Lankapalli AK, Ahmed N (2014) Genome dynamics and evolution of *Salmonella typhi* strains from the typhoid-endemic zones. Sci Rep 4(December):7457

Bakshi U, Sarkar M, Paul S, Dutta C (2016) Assessment of virulence potential of uncharacterized *Enterococcus faecalis* strains using pan genomic approach - identification of pathogen-specific and habitat-specific genes. Sci Rep 6(December):38648

Baltrus DA, Nishimura MT, Romanchuk A, Chang JH, Mukhtar MS, Cherkis K, Roach J, Grant SR, Jones CD, Dangl JL (2011) Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates. PLoS Pathog 7(7):e1002132

Bansal K, Midha S, Kumar S, Patil PB (2017) Ecological and evolutionary insights into *Xanthomonas citri* pathovar diversity. Appl Environ Microbiol 83(9):e02993. https://doi.org/10.1128/AEM.02993-16

Baraúna RA, Ramos RTJ, Veras AAO, Pinheiro KC, Benevides LJ, Viana MVC, Guimarães LC et al (2017) Assessing the genotypic differences between strains of *Corynebacterium* Pseudotuberculosis Biovar Equi through comparative genomics. PLoS One 12(1):e0170676

Bazinet AL (2017) Pan-genome and phylogeny of *Bacillus cereus* Sensu Lato. BMC Evol Biol 17 (1):176

Benevides L, Burman S, Martin R, Robert V, Thomas M, Miquel S, Chain F et al (2017) New insights into the diversity of the genus *Faecalibacterium*. Front Microbiol 8(September):1790

Bergsveinson J, Ziola B (2017) Comparative genomic and plasmid analysis of beer-spoiling and non- beer-spoiling *Lactobacillus brevis* isolates. Can J Microbiol 63(12):970–983

Bhardwaj T, Somvanshi P (2017) Pan-genome analysis of *Clostridium botulinum* reveals unique targets for drug development. Gene 623(August):48–62

Bhattacharyya C, Bakshi U, Mallick I, Mukherji S, Bera B, Ghosh A (2017) Genome-guided insights into the plant growth promotion capabilities of the physiologically versatile *Bacillus aryabhattai* strain AB211. Front Microbiol 8(March):411

Boissy R, Ahmed A, Janto B, Earl J, Hall BG, Hogg JS, Pusch GD et al (2011) Comparative supragenomic analyses among the pathogens *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Haemophilus influenzae* using a modification of the finite supragenome model. BMC Genomics 12(April):187

Bolotin E, Hershberg R (2015) Gene loss dominates as a source of genetic variation within clonal pathogenic bacterial species. Genome Biol Evol 7(8):2173–2187

Borneman AR, McCarthy JM, Chambers PJ, Bartowsky EJ (2012) Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic diversity in industrially-relevant pathways. BMC Genomics 13(August):373

Bottacini F, Motherway MO'C, Kuczynski J, O'Connell KJ, Serafini F, Duranti S, Milani C et al (2014) Comparative genomics of the *Bifidobacterium breve* taxon. BMC Genomics 15 (March):170

Breurec S, Criscuolo A, Diancourt L, Rendueles O, Vandenbogaert M, Passet V, Caro V, Rocha EPC, Touchon M, Brisse S (2016) Genomic epidemiology and global diversity of the emerging bacterial pathogen *Elizabethkingia anophelis*. Sci Rep 6(July):30379

Brito PH, Chevreux B, Serra CR, Schyns G, Henriques AO, Pereira-Leal JB (2018) Genetic competence drives genome diversity in *Bacillus subtilis*. Genome Biol Evol 10(1):108–124

Broadbent JR, Neeno-Eckwall EC, Stahl B, Tandee K, Cai H, Morovic W, Horvath P et al (2012) Analysis of the *Lactobacillus casei* supragenome and its influence in species evolution and lifestyle adaptation. BMC Genomics 13(October):533

Bulagonda EP, Manivannan B, Mahalingam N, Lama M, Chanakya PP, Khamari B, Jadhao S, Vasudevan M, Nagaraja V (2018) Comparative genomic analysis of a naturally competent *Elizabethkingia anophelis* isolated from an eye infection. Sci Rep 8(1):8447

Cao D-M, Qun-Feng L, Li S-B, Wang J-P, Chen Y-L, Huang Y-Q, Bi H-K (2016) Comparative genomics of *H. pylori* and non-*Pylori helicobacter* species to identify new regions associated with its pathogenicity and adaptability. BioMed Res Int:6106029

Cao P, Guo D, Liu J, Jiang Q, Xu Z, Liandong Q (2017) Genome-wide analyses reveal genes subject to positive selection in *Pasteurella multocida*. Front Microbiol 8(May):961

Castillo D, Christiansen RH, Dalsgaard I, Madsen L, Espejo R, Middelboe M (2016) Comparative genome analysis provides insights into the pathogenicity of *Flavobacterium psychrophilum*. PLoS One 11(4):e0152515

Castillo D, Alvise PD, Xu R, Zhang F, Middelboe M, Gram L (2017) Comparative genome analyses of *Vibrio anguillarum* strains reveal a link with pathogenicity traits. mSystems 2(1): e00001. https://doi.org/10.1128/mSystems.00001-17

Castillo D, Pérez-Reytor D, Plaza N, Ramírez-Araya S, Blondel CJ, Corsini G, Bastías R et al (2018) Exploring the genomic traits of non-toxigenic *Vibrio parahaemolyticus* strains isolated in Southern Chile. Front Microbiol 9(February):161

Ceapa C, Davids M, Ritari J, Lambert J, Wels M, Douillard FP, Smokvina T, de Vos WM, Knol J, Kleerebezem M (2016) The variable regions of *Lactobacillus rhamnosus* genomes reveal the dynamic evolution of metabolic and host-adaptation repertoires. Genome Biol Evol 8 (6):1889–1905

Chan AP, Sutton G, DePew J, Krishnakumar R, Choi Y, Huang X-Z, Beck E et al (2015) A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of *Acinetobacter baumannii*. Genome Biol 16(July):143

Chaplin AV, Efimov BA, Smeianov VV, Kafarskaia LI, Pikina AP, Shkoporov AN (2015) Intraspecies genomic diversity and long-term persistence of *Bifidobacterium longum*. PLoS One 10(8):e0135658

Chen C, Wu L, Cao Q, Shao H, Li X, Zhang Y, Wang H, Tan X (2018) Genome comparison of different *Zymomonas mobilis* strains provides insights on conservation of the evolution. PLoS One 13(4):e0195994

Choo SW, Wee WY, Ngeow YF, Mitchell W, Tan JL, Wong GJ, Zhao Y, Xiao J (2014) Genomic reconnaissance of clinical isolates of emerging human pathogen *Mycobacterium abscessus* reveals high evolutionary potential. Sci Rep 4(February):4061

Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, Haley BJ, Taviani E et al (2009) Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. Proc Natl Acad Sci U S A 106(36):15442–15447

Chun BH, Kim KH, Jeon HH, Lee SH, Jeon CO (2017) Pan-genomic and transcriptomic analyses of *Leuconostoc mesenteroides* provide insights into its genomic and metabolic features and roles in Kimchi fermentation. Sci Rep 7(1):11504

Cortés MP, Mendoza SN, Travisany D, Gaete A, Siegel A, Cambiazo V, Maass A (2017) Analysis of *Piscirickettsia salmonis* metabolism using genome-scale reconstruction, modeling, and testing. Front Microbiol 8(December):2462

Coscollá M, Comas I, González-Candelas F (2011) Quantifying nonvertical inheritance in the evolution of *Legionella pneumophila*. Mol Biol Evol 28(2):985–1001

D'Amato F, Eldin C, Georgiades K, Edouard S, Delerce J, Labas N, Raoult D (2015) Loss of TSS1 in hypervirulent *Coxiella burnetii* 175, the causative agent of Q fever in French Guiana. Comp Immunol Microbiol Infect Dis 41(August):35–41

D'Auria G, Jiménez-Hernández N, Peris-Bondia F, Moya A, Latorre A (2010) *Legionella pneumophila* pangenome reveals strain-specific virulence factors. BMC Genomics 11 (March):181

Delmont TO, Murat Eren A (2018) Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. PeerJ 6(January):e4320

De M, Pieter WYC, Rubagotti E, Venter SN, Toth IK, Birch PRJ, Coutinho TA (2014) Analysis of the *Pantoea ananatis* pan-genome reveals factors underlying its ability to colonize and interact with plant, insect and vertebrate hosts. BMC Genomics 15(May):404

Deng X, Phillippy AM, Li Z, Salzberg SL, Zhang W (2010) Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. BMC Genomics 11(September):500

Dias GM, Bidault A, Le Chevalier P, Choquet G, Der Sarkissian C, Orlando L, Medigue C et al (2018) *Vibrio tapetis* displays an original type IV secretion system in strains pathogenic for *Bivalve molluscs*. Front Microbiol 9(February):227

Ding W, Baumdicker F, Neher RA (2018) panX: pan-genome analysis and exploration. Nucleic Acids Res 46(1):e5

Donati C, Luisa Hiller N, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, Oggioni M et al (2010) Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol 11(10):R107

Duchaud E, Rochat T, Habib C, Barbier P, Loux V, Guérin C, Dalsgaard I et al (2018) Genomic diversity and evolution of the fish pathogen *Flavobacterium psychrophilum*. Front Microbiol 9 (February):138

Duranti S, Milani C, Lugli GA, Turroni F, Mancabelli L, Sanchez B, Ferrario C et al (2015) Insights from genomes of representatives of the human gut commensal *Bifidobacterium bifidum*. Environ Microbiol 17(7):2515–2531

Duranti S, Milani C, Lugli GA, Mancabelli L, Turroni F, Ferrario C, Mangifesta M et al (2016) Evaluation of genetic diversity among strains of the human gut commensal *Bifidobacterium adolescentis*. Sci Rep 6(April):23971

Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, Mou S, Achtman M, Lindler LE, Ravel J (2010) Genome sequence of the deep-rooted *Yersinia pestis* strain Angola reveals new insights into the evolution and pangenome of the Plague bacterium. J Bacteriol 192 (6):1685–1699

Fang Y, Li Z, Liu J, Shu C, Wang X, Zhang X, Yu X et al (2011) A pangenomic study of *Bacillus thuringiensis*. J Genet Genomics 38(12):567–576

Fernández-Romero N, Romero-Gómez MP, Mora-Rillo M, Rodríguez-Baño J, López- Cerero L, Pascual Á, Mingorance J (2015) Uncoupling between core genome and virulome in extraintestinal pathogenic *Escherichia coli*. Can J Microbiol 61(9):647–652

Ferrario C, Ricci G, Milani C, Lugli GA, Ventura M, Eraclio G, Borgo F, Fortina MG (2013) *Lactococcus garvieae*: where is it from? A first approach to explore the evolutionary history of this emerging pathogen. PLoS One 8(12):e84796

Fischer W, Windhager L, Rohrer S, Zeiller M, Karnholz A, Hoffmann R, Zimmer R, Haas R (2010) Strain-specific genes of *Helicobacter pylori*: genome evolution driven by a novel type IV secretion system and genomic island transfer. Nucleic Acids Res 38(18):6089–6101

Forgetta V, Oughton MT, Marquis P, Brukner I, Blanchette R, Haub K, Magrini V et al (2011) Fourteen-genome comparison identifies DNA markers for severe-disease-associated strains of *Clostridium difficile*. J Clin Microbiol 49(6):2230–2238

Frese SA, Benson AK, Tannock GW, Loach DM, Kim J, Zhang M, Phaik Lyn O et al (2011) The evolution of host specialization in the vertebrate gut symbiont *Lactobacillus reuteri*. PLoS Genet 7(2):e1001314

Galardini M, Mengoni A, Brilli M, Pini F, Fioravanti A, Lucas S, Lapidus A et al (2011) Exploring the symbiotic pangenome of the nitrogen-fixing bacterium *Sinorhizobium meliloti*. BMC Genomics 12(May):235

Galardini M, Pini F, Bazzicalupo M, Biondi EG, Mengoni A (2013) Replicon-dependent bacterial genome evolution: the case of *Sinorhizobium meliloti*. Genome Biol Evol 5(3):542–558

Garrido-Sanz D, Meier-Kolthoff JP, Göker M, Martín M, Rivilla R, Redondo-Nieto M (2016) Genomic and genetic diversity within the *Pseudomonas fluorescens* complex. PloS One 11(2): e0150183

Ghatak S, Blom J, Das S, Sanjukta R, Puro K, Mawlong M, Shakuntala I et al (2016) Pan-genome analysis of *Aeromonas hydrophila*, *Aeromonas veronii* and *Aeromonas caviae* indicates phylogenomic diversity and greater pathogenic potential for *Aeromonas hydrophila*. *Antonie van Leeuwenhoek* 109(7):945–956

Giampetruzzi A, Saponari M, Loconsole G, Boscia D, Savino VN, Almeida RPP, Zicca S, Landa BB, Chacón-Diaz C, Saldarelli P (2017) Genome-wide analysis provides evidence on the genetic relatedness of the emergent *Xylella fastidiosa* genotype in Italy to isolates from Central America. Phytopathology 107(7):816–827

Gómez-Lunar Z, Hernández-González I, Rodríguez-Torres M-D, Souza V, Olmedo-Álvarez G (2016) Microevolution analysis of *Bacillus coahuilensis* unveils differences in phosphorus acquisition strategies and their regulation. Front Microbiol 7(February):58

Gomila M, Busquets A, Mulet M, García-Valdés E, Lalucat J (2017) Clarification of taxonomic status within the *Pseudomonas syringae* species group based on a phylogenomic analysis. Front Microbiol 8(December):2422

Gordon DM, Geyik S, Clermont O, O'Brien CL, Huang S, Abayasekara C, Rajesh A et al (2017) Fine-scale structure analysis shows epidemic patterns of clonal complex 95, a cosmopolitan *Escherichia coli* lineage responsible for extraintestinal infection. mSphere 2(3):e00168. https://doi.org/10.1128/mSphere.00168-17

Grosso-Becerra M-V, Santos-Medellín C, González-Valdez A, Méndez J-L, Delgado G, Morales-Espinosa R, Servín-González L, Alcaraz L-D, Soberón-Chávez G (2014) *Pseudomonas aeruginosa* clinical and environmental isolates constitute a single population with high phenotypic diversity. BMC Genomics 15(April):318

Guo X, Li S, Zhang J, Feifan W, Li X, Dan W, Zhang M et al (2017) Genome sequencing of 39 *Akkermansia muciniphila* isolates reveals its population structure, genomic and functional diversity, and global distribution in mammalian gut microbiotas. BMC Genomics 18(1):800

Hall AB, Yassour M, Sauk J, Garner A, Jiang X, Arthur T, Lagoudas GK et al (2017) A novel *Ruminococcus gnavus* clade enriched in inflammatory bowel disease patients. Genome Med 9 (1):103

Hassan A, Naz A, Obaid A, Paracha RZ, Naz K, Awan FM, Muhmmad SA, Janjua HA, Ahmad J, Ali A (2016) Pangenome and immuno-proteomics analysis of *Acinetobacter baumannii* strains revealed the core peptide vaccine targets. BMC Genomics 17(1):732

He E-M, Chen C-W, Guo Y, Hsu M-H, Liang Z, Chen H-L, Zhao G-P, Chiu C-H, Zhou Y (2017) The genome of serotype VI *Streptococcus agalactiae* serotype VI and comparative analysis. Gene 597(January):59–65

Hilker R, Munder A, Klockgether J, Losada PM, Chouvarine P, Cramer N, Davenport CF et al (2015) Interclonal gradient of virulence in the *Pseudomonas aeruginosa* pangenome from disease and environment. Environ Microbiol 17(1):29–46

Hiller NL, Janto B, Hogg JS, Boissy R, Susan Y, Powell E, Keefe R et al (2007) Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J Bacteriol 189(22):8186–8195

Hollensteiner J, Poehlein A, Spröer C, Bunk B, Sheppard AE, Rosentstiel P, Schulenburg H, Liesegang H (2017) Complete genome sequence of the nematicidal *Bacillus thuringiensis* MYBT18246. Stand Genomic Sci 12(August):48

Holm KO, Bækkedal C, Söderberg JJ, Haugen P (2018) Complete genome sequences of seven *Vibrio anguillarum* strains as derived from PacBio sequencing. Genome Biol Evol 10 (4):1127–1131

Howell KJ, Weinert LA, Chaudhuri RR, Luan S-L, Peters SE, Corander J, Harris D et al (2014) The use of genome wide association methods to investigate pathogenicity, population structure and serovar in *Haemophilus parasuis*. BMC Genomics 15(December):1179

Howell KJ, Weinert LA, Peters SE, Wang J, Hernandez-Garcia J, Chaudhuri RR, Luan S-L et al (2017) 'Pathotyping' multiplex PCR assay for *Haemophilus parasuis*: a tool for prediction of virulence. J Clin Microbiol 55(9):2617–2628

Huang Y, Kittichotirat W, Mayer MPA, Hall R, Bumgarner R, Chen C (2013) Comparative genomic hybridization and transcriptome analysis with a pan-genome microarray reveal distinctions between JP2 and non-JP2 genotypes of *Aggregatibacter actinomycetemcomitans*. Mol Oral Microbiol 28(1):1–17

Humbert J-F, Barbe V, Latifi A, Gugger M, Calteau A, Coursin T, Lajus A et al (2013) A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*. PLoS One 8(8):e70747

Hurtado R, Carhuaricra D, Soares S, Viana MVC, Azevedo V, Maturrano L, Aburjaile F (2018) Pan-genomic approach shows insight of genetic divergence and pathogenic-adaptation of *Pasteurella multocida*. Gene 670(September):193–206

Imperi F, Antunes LCS, Blom J, Villa L, Iacono M, Visca P, Carattoli A (2011) The genomics of *Acinetobacter baumannii*: insights into genome plasticity, antimicrobial resistance and pathogenicity. IUBMB Life 63(12):1068–1074

Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW, Friis C (2011) The *Salmonella enterica* pan-genome. Microb Ecol 62(3):487–504

Janvilisri T, Scaria J, Thompson AD, Nicholson A, Limbago BM, Arroyo LG, Glenn Songer J, Gröhn YT, Chang Y-F (2009) Microarray identification of *Clostridium difficile* core components and divergent regions associated with host origin. J Bacteriol 191(12):3881–3891

Jeong D-W, Heo S, Ryu S, Blom J, Lee J-H (2017) Genomic insights into the virulence and salt tolerance of *Staphylococcus equorum*. Sci Rep 7(1):5383

Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD (2011) Interplay of recombination and selection in the genomes of *Chlamydia trachomatis*. Biol Direct 6(May):28

Joseph SJ, Cox D, Wolff B, Morrison SS, Kozak-Muiznieks NA, Frace M, Didelot X et al (2016) Dynamics of genome change among Legionella species. Sci Rep 6(September):33442

Kaas RS, Friis C, Ussery DW, Aarestrup FM (2012) Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. BMC Genomics 13(October):577

Kadam A, Janto B, Eutsey R, Earl JP, Powell E, Dahlgren ME, Hu FZ, Ehrlich GD, Luisa Hiller N (2015) *Streptococcus pneumoniae* supragenome hybridization arrays for profiling of genetic content and gene expression. Curr Protoc Microbiol 36(February):9D-4

Kant R, Rintahaka J, Yu X, Sigvart-Mattila P, Paulin L, Mecklin J-P, Saarela M, Palva A, von Ossowski I (2014) A comparative pan-genome perspective of niche-adaptable cell-surface protein phenotypes in *Lactobacillus rhamnosus*. PLoS One 9(7):e102762

Karlsen C, Hjerde E, Klemetsen T, Willassen NP (2017) Pan genome and CRISPR analyses of the bacterial fish pathogen *Moritella viscosa*. BMC Genomics 18(1):313

Kawasaki M, Delamare-Deboutteville J, Bowater RO, Walker MJ, Beatson S, Ben Zakour NL, Barnes AC (2018) Microevolution of aquatic *Streptococcus agalactiae* ST-261 from Australia indicates dissemination via imported tilapia and ongoing adaptation to marine hosts or environment. Appl Environ Microbiol 84(16):e00859. https://doi.org/10.1128/AEM.00859-18

Kayansamruaj P, Pirarat N, Kondo H, Hirono I, Rodkhum C (2015) Genomic comparison between pathogenic *Streptococcus agalactiae* isolated from Nile Tilapia in Thailand and fish-derived ST7 strains. Infect Genet Evol 36(December):307–314

Kayansamruaj P, Dong HT, Hirono I, Kondo H, Senapin S, Rodkhum C (2017) Comparative genome analysis of fish pathogen *Flavobacterium columnare* reveals extensive sequence diversity within the species. Infect Genet Evol 54(October):7–17

Kelleher P, Bottacini F, Mahony J, Kilcawley KN, van Sinderen D (2017) Comparative and functional genomics of the *Lactococcus lactis* taxon; insights into evolution and niche adaptation. BMC Genomics 18(1):267

Kim EB, Marco ML (2014) Nonclinical and clinical *Enterococcus faecium* strains, but not *Enterococcus faecalis* strains, have distinct structural and functional genomic features. Appl Environ Microbiol 80(1):154–165

Kim Y, Koh I, Mi YL, Chung W-H, Rho M (2017) Pan-genome analysis of Bacillus for microbiome profiling. Sci Rep 7(1):10984

Kirk KF, Méric G, Nielsen HL, Pascoe B, Sheppard SK, Thorlacius-Ussing O, Nielsen H (2018) Molecular epidemiology and comparative genomics of campylobacter concisus strains from saliva, faeces and gut mucosal biopsies in inflammatory bowel disease. Sci Rep 8(1):1902

Kittichotirat W, Bumgarner RE, Asikainen S, Chen C (2011) Identification of the pangenome and its components in 14 distinct *Aggregatibacter actinomycetemcomitans* strains by comparative genomic analysis. PLoS One 6(7):e22420

Kiu R, Caim S, Alexander S, Pachori P, Hall LJ (2017) Probing genomic aspects of the multi-host pathogen *Clostridium perfringens* reveals significant pangenome diversity, and a diverse array of virulence factors. Front Microbiol 8(December):2485

Klockgether J, Cramer N, Wiehlmann L, Davenport CF, Tümmler B (2011) *Pseudomonas aeruginosa* genomic structure and diversity. Front Microbiol 2(July):150

Knight DR, Squire MM, Collins DA, Riley TV (2016) Genome analysis of *Clostridium difficile* PCR ribotype 014 lineage in Australian pigs and humans reveals a diverse genetic repertoire and signatures of long-range interspecies transmission. Front Microbiol 7:2138

Koton Y, Gordon M, Chalifa-Caspi V, Bisharat N (2014) Comparative genomic analysis of clinical and environmental *Vibrio vulnificus* isolates revealed biotype 3 evolutionary relationships. Front Microbiol 5:803

Kubasova T, Cejkova D, Matiasovicova J, Sekelova Z, Polansky O, Medvecky M, Rychlik I, Juricova H (2016) Antibiotic resistance, core-genome and protein expression in IncHI1 plasmids in *Salmonella typhimurium*. Genome Biol Evol 8(6):1661–1671

Kuenne C, Billion A, Mraheil MA, Strittmatter A, Daniel R, Goesmann A, Barbuddhe S, Hain T, Chakraborty T (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(January):47

Kunadu P-H, Angela MH, Miller EL, Grant AJ (2018) Microbiological quality and antimicrobial resistance characterization of *Salmonella* spp. in fresh milk value chains in Ghana. Int J Food Microbiol 277(July):41–49

Lacey JA, Allnutt TR, Vezina B, Van TTH, Stent T, Han X, Rood JI et al (2018) Whole genome analysis reveals the diversity and evolutionary relationships between necrotic enteritis-causing strains of *Clostridium perfringens*. BMC Genomics 19(1):379

Laing C, Villegas A, Taboada EN, Kropinski A, Thomas JE, Gannon VPJ (2011) Identification of *Salmonella enterica* species- and subgroup-specific genomic regions using Panseq 2.0. Infect Genet Evol 11(8):2151–2161

Laing CR, Whiteside MD, 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(July):1345

Lebreton F, van Schaik W, McGuire AM, Godfrey P, Griggs A, Mazumdar V, Corander J et al (2013) Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. mBio 4(4):e00534. https://doi.org/10.1128/mBio.00534-13

Lee J-Y, Han GG, Choi J, Jin G-D, Kang S-K, Chae BJ, Kim EB, Choi Y-J (2017a) Pan-genomic approaches in *Lactobacillus reuteri* as a porcine probiotic: investigation of host adaptation and antipathogenic activity. Microb Ecol 74(3):709–721

Lee J-Y, Han GG, Kim EB, Choi Y-J (2017b) Comparative genomics of *Lactobacillus salivarius* strains focusing on their host adaptation. Microbiol Res 205(December):48–58

Leekitcharoenphon P, Lukjancenko O, Friis C, Aarestrup FM, Ussery DW (2012) Genomic variation in *Salmonella enterica* core genes for epidemiological typing. BMC Genomics 13 (March):88

Liang W, Zhao Y, Chen C, Cui X, Yu J, Xiao J, Kan B (2012) Pan-genomic analysis provides insights into the genomic variation and evolution of *Salmonella paratyphi* A. PLoS One 7(9): e45346

Li G, Shen M, Le S, Tan Y, Li M, Zhao X, Shen W et al (2016) Genomic analyses of multidrug resistant *Pseudomonas aeruginosa* PA1 resequenced by single-molecule real-time sequencing. Biosci Rep 36(6):e00418. https://doi.org/10.1042/BSR20160282

Lira F, Berg G, Martínez JL (2017) Double-face meets the bacterial world: the opportunistic pathogen *Stenotrophomonas maltophilia*. Front Microbiol 8(November):2190

Liu W, Fang L, Li M, Li S, Guo S, Luo R, Feng Z et al (2012) Comparative genomics of mycoplasma: analysis of conserved essential genes and diversity of the pan-genome. PLoS One 7(4):e35698

Liu L, Zhu W, Cao Z, Biao X, Wang G, Luo M (2015) High correlation between genotypes and phenotypes of environmental bacteria *Comamonas testosteroni* strains. BMC Genomics 16 (February):110

Liu G, Zhang W, Chengping L (2013a) Comparative genomics analysis of *Streptococcus agalactiae* reveals that isolates from cultured Tilapia in China are closely related to the human strain A909. BMC Genomics 14(November):775

Liu W-Y, Wong C-F, Chung KM-K, Jiang J-W, Leung FC-C (2013b) Comparative genome analysis of Enterobacter cloacae. PLoS One 8(9):e74487

Liu Y-Y, Chen C-C, Chiou C-S (2016a) Construction of a pan-genome allele database of *Salmonella enterica* serovar enteritidis for molecular subtyping and disease cluster identification. Front Microbiol 7(December):2010

Liu Y-Y, Chiou C-S, Chen C-C (2016b) PGAdb-builder: a web service tool for creating pan-genome allele database for molecular fine typing. Sci Rep 6(November):36213

Loper JE, Hassan KA, Mavrodi DV, Davis EW 2nd, Lim CK, Shaffer BT, Elbourne LDH et al (2012) Comparative genomics of plant-associated *Pseudomonas* spp.: insights into diversity and inheritance of traits involved in multitrophic interactions. PLoS Genet 8(7):e1002784

López-Hermoso C, de la Haba RR, Sánchez-Porro C, Ventosa A (2018) *Salinivibrio kushneri* sp. nov., a moderately halophilic bacterium isolated from salterns. Syst Appl Microbiol 41 (3):159–166

Lukjancenko O, Wassenaar TM, Ussery DW (2010) Comparison of 61 sequenced *Escherichia Coli* genomes. Microb Ecol 60(4):708–720

Lu W, Wise MJ, Tay CY, Windsor HM, Marshall BJ, Peacock C, Perkins T (2014) Comparative analysis of the full genome of *Helicobacter pylori* isolate Sahul64 identifies genes of high divergence. J Bacteriol 196(5):1073–1083

Malmstrom RR, Rodrigue S, Huang KH, Kelly L, Kern SE, Thompson A, Roggensack S, Berube PM, Henn MR, Chisholm SW (2013) Ecology of uncultured *Prochlorococcus* clades revealed through single-cell genomics and biogeographic analysis. ISME J 7(1):184–198

Mann RA, Smits THM, Bühlmann A, Blom J, Goesmann A, Frey JE, Plummer KM et al (2013) Comparative genomics of 12 strains of Erwinia Amylovora identifies a pan-genome with a large conserved core. PLoS One 8(2):e55644

Manzano-Marín A, Lamelas A, Moya A, Latorre A (2012) Comparative genomics of *Serratia* spp.: two paths towards endosymbiotic life. PLoS One 7(10):e47274

Meng P, Lu C, Zhang Q, Lin J, Chen F (2017) Exploring the genomic diversity and cariogenic differences of *Streptococcus mutans* strains through pan-genome and comparative genome analysis. Curr Microbiol 74(10):1200–1209

Méric G, Yahara K, Mageiros L, Pascoe B, Maiden MCJ, Jolley KA, Sheppard SK (2014) A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic campylobacter. PLoS One 9(3):e92798

Méric G, Miragaia M, de Been M, Yahara K, Pascoe B, Mageiros L, Mikhail J et al (2015) Ecological overlap and horizontal gene transfer in *Staphylococcus aureus* and *Staphylococcus epidermidis*. Genome Biol Evol 7(5):1313–1328

Méric G, Mageiros L, Pascoe B, Woodcock DJ, Mourkas E, Lamble S, Bowden R, Jolley KA, Raymond B, Sheppard SK (2018) Lineage-specific plasmid acquisition and the evolution of specialized pathogens in *Bacillus thuringiensis* and the *Bacillus cereus* group. Mol Ecol 27 (7):1524–1540

Milani C, Duranti S, Lugli GA, Bottacini F, Strati F, Arioli S, Foroni E, Turroni F, van Sinderen D, Ventura M (2013) Comparative genomics of *Bifidobacterium animalis* subsp. lactis reveals a strict monophyletic bifidobacterial taxon. Appl Environ Microbiol 79(14):4304–4315

Miyauchi E, Toh H, Nakano A, Tanabe S, Morita H (2012) Comparative genomic analysis of *Lactococcus garvieae* strains isolated from different sources reveals candidate virulence genes. Int J Microbiol 2012(May):728276

Mongodin EF, Casjens SR, Bruno JF, Yun X, Drabek EF, Riley DR, Cantarel BL et al (2013) Inter- and intra-specific pan-genomes of *Borrelia burgdorferi* Sensu Lato: genome stability and adaptive radiation. BMC Genomics 14(October):693

Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, Feist AM, Palsson BØ (2013) Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. Proc Natl Acad Sci U S A 110 (50):20338–20343

Mosquera-Rendón J, Rada-Bravo AM, Cárdenas-Brito S, Corredor M, Restrepo-Pineda E, Benítez-Páez A (2016) Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species. BMC Genomics 17(January):45

Murillo T, Ramírez-Vargas G, Riedel T, Overmann J, Andersen JM, Guzmán-Verri C, Chaves-Olarte E, Rodríguez C (2018) Two groups of cocirculating, epidemic clostridiodes difficile strains microdiversify through different mechanisms. Genome Biol Evol 10(3):982–998

Nguyen TL, Kim D-H (2018) Genome-wide comparison reveals a probiotic strain *Lactococcus lactis* WFLU12 isolated from the gastrointestinal tract of olive flounder (*Paralichthys olivaceus*) harboring genes supporting probiotic action. Mar Drugs 16(5):140. https://doi.org/10.3390/md16050140

Nourdin-Galindo G, Sánchez P, Molina CF, Espinoza-Rojas DA, Oliver C, Ruiz P, Vargas-Chacoff L et al (2017) Comparative pan-genome analysis of *Piscirickettsia salmonis* reveals genomic divergences within genogroups. Front Cell Infect Microbiol 7(October):459

O'Callaghan A, Bottacini F, O'Connell Motherway M, van Sinderen D (2015) Pangenome analysis of *Bifidobacterium longum* and site-directed mutagenesis through by-pass of restriction-modification systems. BMC Genomics 16(October):832

Ogunremi D, Devenish J, Amoako K, Kelly H, Dupras AA, Belanger S, Wang LR (2014) High resolution assembly and characterization of genomes of Canadian isolates of *Salmonella enteritidis*. BMC Genomics 15(August):713

Ojala T, Kankainen M, Castro J, Cerca N, Edelman S, Westerlund-Wikström B, Paulin L, Holm L, Auvinen P (2014) Comparative genomics of *Lactobacillus crispatus* suggests novel mechanisms for the competitive exclusion of *Gardnerella vaginalis*. BMC Genomics 15 (December):1070

Okura M, Nozawa T, Watanabe T, Murase K, Nakagawa I, Takamatsu D, Osaki M et al (2017) A locus encoding variable defence systems against invading DNA identified in *Streptococcus suis*. Genome Biol Evol 9(4):1000. https://doi.org/10.1093/gbe/evx062

Orata FD, Kirchberger PC, Méheust R, Barlow EJ, Tarr CL, Boucher Y (2015) The dynamics of genetic interactions between *Vibrio metoecus* and *Vibrio cholerae*, two close relatives co-occurring in the environment. Genome Biol Evol 7(10):2941–2954

Otchere ID, Harris SR, Busso SL, Asante-Poku A, Osei-Wusu S, Koram K, Parkhill J, Gagneux S, Yeboah-Manu D (2016) The first population structure and comparative genomics analysis of mycobacterium africanum strains from Ghana reveals higher diversity of lineage. Int J Mycobact 5(Suppl 1):S80–S81

Palmer SR, Miller JH, Abranches J, Zeng L, Lefebure T, Richards VP, Lemos JA, Stanhope MJ, Burne RA (2013) Phenotypic heterogeneity of genomically-diverse isolates of *Streptococcus mutans*. PLoS One 8(4):e61358

Periwal V, Patowary A, Vellarikkal SK, Gupta A, Singh M, Mittal A, Jeyapaul S et al (2015) Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of *Mycobacterium tuberculosis* pangenome. PLoS One 10(4):e0122979

Qin X, Galloway-Peña JR, Sillanpaa J, Roh JH, Nallapareddy SR, Chowdhury S, Bourgogne A et al (2012) Complete genome sequence of *Enterococcus faecium* strain TX16 and comparative genomic analysis of *Enterococcus faecium* genomes. BMC Microbiol 12(July):135

Remenant B, Coupat-Goutaland B, Guidot A, Cellier G, Wicker E, Allen C, Fegan M et al (2010) Genomes of three tomato pathogens within the *Ralstonia solanacearum* species complex reveal significant evolutionary divergence. BMC Genomics 11(June):379

Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ (2009) Biogeography of the *Sulfolobus islandicus* pan-genome. Proc Natl Acad Sci U S A 106(21):8605–8610

Roisin S, Gaudin C, De Mendonça R, Bellon J, Van Vaerenbergh K, De Bruyne K, Byl B, Pouseele H, Denis O, Supply P (2016) Pan-genome multilocus sequence typing and outbreak-specific reference-based single nucleotide polymorphism analysis to resolve two concurrent *Staphylococcus aureus* outbreaks in neonatal services. Clin Microbiol Infect 22(6):520–526

Rouleau FD, Vincent AT, Charette SJ (2018) Genomic and phenotypic characterization of an atypical *Aeromonas salmonicida* strain isolated from a lumpfish and producing unusual granular structures. J Fish Dis 41(4):673–681

Rouli L, MBengue M, Robert C, Ndiaye M, La Scola B, Raoult D (2014) Genomic analysis of three African strains of *Bacillus anthracis* demonstrates that they are part of the clonal expansion of an exclusively pathogenic bacterium. New Microb New Infect 2(6):161–169

Sangal V, Blom J, Sutcliffe IC, von Hunolstein C, Burkovski A, Hoskisson PA (2015) Adherence and invasive properties of *Corynebacterium diphtheriae* strains correlates with the predicted membrane-associated and secreted proteome. BMC Genomics 16(October):765

Sassi M, Drancourt M (2014) Genome analysis reveals three genomospecies in *Mycobacterium abscessus*. BMC Genomics 15(May):359

Scaria J, Ponnala L, Janvilisri T, Yan W, Mueller LA, Chang Y-F (2010) Analysis of ultra low genome conservation in *Clostridium difficile*. PLoS One 5(12):e15147

Sela U, Euler CW, Correa da Rosa J, Fischetti VA (2018) Strains of bacterial species induce a greatly varied acute adaptive immune response: the contribution of the accessory genome. PLoS Pathog 14(1):e1006726

Shariati J V, Malboobi MA, Tabrizi Z, Tavakol E, Owilia P, Safari M (2017) Comprehensive genomic analysis of a plant growth-promoting rhizobacterium pantoea agglomerans strain P5. Sci Rep 7(1):15610

Sharma PK, Jilagamazhi F, Zhang X, Fristensky B, Sparling R, Levin DB (2014) Genome features of *Pseudomonas putida* LS46, a novel polyhydroxyalkanoate producer and its comparison with other P. putida strains. AMB Express 4(May):37

Silby MW, Cerdeño-Tárraga AM, Vernikos GS, Giddens SR, Jackson RW, Preston GM, Zhang X-X et al (2009) Genomic and genetic analyses of diversity and plant interactions of *Pseudomonas fluorescens*. Genome Biol 10(5):R51

Smokvina T, Wels M, Polka J, Chervaux C, Brisse S, Boekhorst J, Vlieg JET v H, Siezen RJ (2013) *Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity. PLoS One 8(7):e68731

Snipen L, Almøy T, Ussery DW (2009) Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 10(August):385

Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A et al (2013) The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the *Biovar ovis* and equi strains. PLoS One 8(1):e53818

Song L, Wang W, Conrads G, Rheinberg A, Sztajer H, Reck M, Wagner-Döbler I, Ping Zeng A (2013) Genetic variability of *Mutans streptococci* revealed by wide whole-genome sequencing. BMC Genomics 14(June):430

Spring-Pearson SM, Stone JK, Doyle A, Allender CJ, Okinaka RT, Mayo M, Broomall SM et al (2015) Pangenome analysis of *Burkholderia pseudomallei*: genome evolution preserves gene order despite high recombination rates. PLoS One 10(10):e0140274

Stabler RA, Gerding DN, Songer JG, Drudy D, Brazier JS, Trinh HT, Witney AA, Hinds J, Wren BW (2006) Comparative phylogenomics of *Clostridium difficile* reveals clade specificity and microevolution of hypervirulent strains. J Bacteriol 188(20):7297–7305

Stanborough T, Fegan N, Powell SM, Singh T, Tamplin M, Scott Chandry P (2018) Genomic and metabolic characterization of spoilage-associated Pseudomonas species. Int J Food Microbiol 268(March):61–72

Sternes PR, Borneman AR (2016) Consensus pan-genome assembly of the specialised wine bacterium *Oenococcus oeni*. BMC Genomics 17(April):308

Stice SP, Stumpf SD, Gitaitis RD, Kvitko BH, Dutta B (2018) *Pantoea ananatis* genetic diversity analysis reveals limited genomic diversity as well as accessory genes correlated with onion pathogenicity. Front Microbiol 9(February):184

Sultanov RI, Arapidi GP, Vinogradova SV, Govorun VM, Luster DG, Ignatov AN (2016) Comprehensive analysis of draft genomes of two closely related *Pseudomonas syringae* phylogroup 2b strains infecting mono- and dicotyledon host plants. BMC Genomics 17(Suppl 14):1010

Sun S, Xiao J, Zhang H, Zhang Z (2016) Pangenome evidence for higher codon usage bias and stronger translational selection in core genes of *Escherichia coli*. Front Microbiol 7 (August):1180

Sváb D, Bálint B, Maróti G, Tóth I (2016) Cytolethal distending toxin producing *Escherichia coli* O157:H43 strain T22 represents a novel evolutionary lineage within the O157 serogroup. Infect Genet Evol 46(December):110–117

Thépault A, Méric G, Rivoal K, Pascoe B, Mageiros L, Touzain F, Rose V, Béven V, Chemaly M, Sheppard SK (2017) Genome-wide identification of host- segregating epidemiological markers for source attribution in *Campylobacter jejuni*. Appl Environ Microbiol 83(7):e03085. https://doi.org/10.1128/AEM.03085-16

Timms VJ, Rockett R, Bachmann NL, Martinez E, Wang Q, Chen SC-A, Jeoffreys N et al (2018) Genome sequencing links persistent outbreak of legionellosis in Sydney (New South Wales, Australia) to an emerging clone of *Legionella pneumophila* sequence type 211. Appl Environ Microbiol 84(5):e02020. https://doi.org/10.1128/AEM.02020-17

Tomida S, Nguyen L, Chiu B-H, Liu J, Sodergren E, Weinstock GM, Li H (2013) Pan-genome and comparative genome analyses of *Propionibacterium acnes* reveal its genomic diversity in the healthy and diseased human skin microbiome. mBio 4(3):e00003

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E et al (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 5(1):e1000344

Trost E, Blom J, Soares S d C, Huang I-H, Al-Dilaimi A, Schröder J, Jaenicke S et al (2012) Pangenomic study of *Corynebacterium diphtheriae* that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. J Bacteriol 194(12):3199–3215

Uchiyama I, Albritton J, Fukuyo M, Kojima KK, Yahara K, Kobayashi I (2016) A novel approach to *Helicobacter pylori* pan-genome analysis for identification of genomic islands. PLoS One 11 (8):e0159419

Udaondo Z, Molina L, Segura A, Duque E, Ramos JL (2016) Analysis of the core genome and pangenome of *Pseudomonas putida*. Environ Microbiol 18(10):3268–3283 van Schaik W, Top J, Riley DR, Boekhorst J, Vrijenhoek JEP, Schapendonk CME, Hendrickx APA et al (2010) Pyrosequencing-based comparative genome analysis of the nosocomial pathogen *Enterococcus faecium* and identification of a large transferable pathogenicity island. BMC Genomics 11(April):239

Vesth T, Wassenaar TM, Hallin PF, Snipen L, Lagesen K, Ussery DW (2010) On the origins of a Vibrio species. Microb Ecol 59(1):1–13

Viver T, Orellana L, González-Torres P, Díaz S, Urdiain M, Farías ME, Benes V et al (2018) Genomic comparison between members of the Salinibacteraceae family, and description of a new species of Salinibacter (*Salinibacter altiplanensis* sp. nov.) isolated from high altitude hypersaline environments of the Argentinian Altiplano. Syst Appl Microbiol 41(3):198–212

Wang J, Haapalainen M, Schott T, Thompson SM, Smith GR, Nissinen AI, Pirhonen M (2017) Genomic sequence of 'candidatus liberibacter solanacearum' haplotype C and its comparison with haplotype A and B genomes. PLoS One 12(2):e0171531

Wegmann U, MacKenzie DA, Zheng J, Goesmann A, Roos S, Swarbreck D, Walter J, Crossman LC, Juge N (2015) The pan-genome of *Lactobacillus reuteri* strains originating from the pig gastrointestinal tract. BMC Genomics 16(December):1023

Weller-Stuart T, De Maayer P, Coutinho T (2017) Pantoea ananatis: genomic insights into a versatile pathogen. Mol Plant Pathol 18(9):1191–1198

Wilkinson DA, O'Donnell AJ, Akhter RN, Fayaz A, Mack HJ, Rogers LE, Biggs PJ, French NP, Midwinter AC (2018) Updating the genomic taxonomy and epidemiology of *Campylobacter hyointestinalis*. Sci Rep 8(1):2393

Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW (2007) Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. Genome Biol 8(12):R267

Williams TM, Loman NJ, Ebruke C, Musher DM, Adegbola RA, Pallen MJ, Weinstock GM, Antonio M (2012) Genome analysis of a highly virulent serotype 1 strain of *Streptococcus pneumoniae* from West Africa. PLoS One 7(10):e26742

Xu Z, Chen X, Li L, Li T, Wang S, Chen H, Zhou R (2010) Comparative genomic characterization of *Actinobacillus pleuropneumoniae*. J Bacteriol 192(21):5625–5636

Yang J, Yang S (2017) Comparative analysis of corynebacterium glutamicum genomes: a new perspective for the industrial production of amino acids. BMC Genomics 18(Suppl 1):940

Yu G, Wang XC, Tian WH, Shi JC, Wang B, Ye Q, Dong SG, Zeng M, Wang JZ (2015) Genomic diversity and evolution of *Bacillus subtilis*. Biomed Environ Sci 28(8):620–625

Yu D, Yin Z, Li B, Jin Y, Ren H, Zhou J, Zhou W, Liang L, Yue J (2016) Gene flow, recombination, and positive selection in *Stenotrophomonas maltophilia*: mechanisms underlying the diversity of the widespread opportunistic pathogen. Genome 59(12):1063–1075

Yu J, Zhao J, Song Y, Zhang J, Yu Z, Zhang H, Sun Z (2018) Comparative genomics of the herbivore gut symbiont *Lactobacillus reuteri* reveals genetic diversity and lifestyle adaptation. Front Microbiol 9(June):1151

Yue M, Rankin SC, Blanchet RT, Nulton JD, Edwards RA, Schifferli DM (2012) Diversification of the *Salmonella fimbriae*: a model of macro- and microevolution. PLoS One 7(6):e38596

Zhang A, Yang M, Hu P, Wu J, Chen B, Hua Y, Yu J, Chen H, Xiao J, Jin M (2011a) Comparative genomic analysis of *Streptococcus suis* reveals significant genomic diversity among different serotypes. BMC Genomics 12(October):523

Zhang J, van Aartsen JJ, Jiang X, Shao Y, Tai C, He X, Tan Z et al (2011b) Expansion of the known *Klebsiella pneumoniae* species gene pool by characterization of novel alien DNA islands integrated into tmRNA gene sites. J Microbiol Methods 84(2):283–289

Zhang D-F, Zhi X-Y, Zhang J, Paoli GC, Cui Y, Shi C, Shi X (2017) Preliminary comparative genomics revealed pathogenic potential and international spread of *Staphylococcus argenteus*. BMC Genomics 18(1):808

Zhang X, Liu X, Yang F, Chen L (2018) Pan-genome analysis links the hereditary variation of *Leptospirillum ferriphilum* with its evolutionary adaptation. Front Microbiol 9(March):577

Zhao Y, Sun C, Zhao D, Zhang Y, You Y, Jia X, Yang J et al (2018) PGAP-X: extension on pan-genome analysis pipeline. BMC Genomics 19(Suppl 1):36

Zhou Z, McCann A, Litrup E, Murphy R, Cormican M, Fanning S, Brown D, Guttman DS, Brisse S, Achtman M (2013) Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* serovar agona. PLoS Genet 9(4):e1003471

Zhu Ge X, Jiang J, Pan Z, Hu L, Wang S, Wang H, Leung FC, Dai J, Fan H (2014) Comparative genomic analysis shows that avian pathogenic *Escherichia coli* isolate IMT5155 (O2:K1:H5; ST Complex 95, ST140) shares close relationship with ST95 APEC O1:K1 and human ExPEC O18:K1 strains. PLoS One 9(11):e112048

A Review of Pangenome Tools and Recent Studies

G. S. Vernikos

Abstract With the advance of sequencing technologies, the landscape of genomic analysis has been transformed, by moving from single strain to species (or even higher taxa)-wide genomic resolution, toward the direction of capturing the "totality" of life diversity; from this scientific advance and curiosity, the concept of "pangenome" was born. Herein we will review, from practical and technical implementation, existing projects of pangenome analysis, with the aim of providing the reader with a snapshot of useful tools should they need to embark on such a pangenomic journey.

Keywords Pangenome · Whole-genome · Exhaustive search · Subsampling · Regression function · Command line · Web-interface · Bayesian · Hidden Markov Models · Clustering · ORF alignment similarity · Combinatorial approach · Ortholog clusters · Reference pangenome · Finite supragenome model · Binomial mixture model · Infinitely many genes model · Gene presence/absence frequency

1 Introduction

Almost 15 years ago, Tettelin et al. (2005) conceived the concept of pangenome, in an attempt to describe and model the genomic totality of a taxa (species, serovar, phylum, kingdom, etc.) of interest. Since then the nomenclature of this concept became fairly wide to accommodate words like pangenome, core and dispensable genes, strain-specific genes (Medini et al. 2005; Tettelin et al. 2005), supragenome, distributed and unique genes (Lapierre and Gogarten 2009), and flexible regions (Rodriguez-Valera and Ussery 2012). Simply put, using the original definition, the core-genome describes the set of sequences shared by all members of the taxa of interest, the dispensable genome captures a subset of sequences shared by some

G. S. Vernikos (✉)
GlaxoSmithKline, Medical Affairs Department, Athens, Greece
e-mail: georgios.x.vernikos@gsk.com

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_4

members of the group (dictating the diversity of the group: alternative biochemical pathways, niche adaptation, antibiotic resistance, etc.) while the pangenome is simply the union of core and dispensable genomes (describing the totality of taxa at the level of sequence datasets).

The exponential growth of genomic databases started in 1995 with *Haemophilus influenzae* being the first complete genome project (Fleischmann et al. 1995). Today, as of August 2018, 110,660 complete whole-genome sequencing projects—of which 87% are bacteria—and 15,066 finished whole-genome sequencing projects (Mukherjee et al. 2017) are available in the public domain. These fueled the interest of many researchers to carry out pangenome analysis at every conceivable phylogenetic resolution level (Table 1), exploiting various modeling frameworks, assumptions, and underlying homology search engines.

A pivotal work in terms of phylogenetic resolution was carried out by Lapierre and Gogarten (2009), showing that on average in the largest bacterium group analyzed so far, the core gene set accounts only for 8% of the pangenome.

The pangenome concept can be implemented either in reverse or in forward-thinking approaches; in the first case, we are interested to capture the genomic diversity of the group of interest, while in the second case we are more interested in exploring and predicting from a pragmatic perspective what is the minimum number of genome sequences required to capture the totality of the group. Obviously, limited or sparse datasets might lead to erroneous conclusions; therefore, it was recommended (Vernikos et al. 2015) that the minimum number of genomes to analyze be at least five.

The lifestyle of the species of interest is one of the parameters strongly dictating the distribution shape of the pangenome; for example, if by recurring addition of group members, the pangenome continues to grow, we are analyzing an open pangenome (such examples include human pathogens and environmental bacteria) (Hiller et al. 2007; Tettelin et al. 2008). On the other hand, if the group complexity is exhausted very fast even from the analyses of a handful of group members then we are dealing with a closed pangenome whereby we only need few representatives to describe the totality of the sequence variability.

2 Technical Implementation

In pangenome analysis, the sequence unit for the modeling can be anything from ORFs, genes, clusters of orthologous groups COGs (Tatusov et al. 1997), coding sequences (CDS), proteins, arbitrary sequence chunks, concatenated gene or protein entities, etc.

Practical aspects of consideration that directly influence the validity of the conclusions drawn, include how quickly is expected a pangenome to grow and reach a plateau (open or close pangenome), the parameters that determine in the search engine the orthologous sequences and thereby directly affect the pool of core and dispensable sequence entities, the mathematical model and the applied

Table 1 Examples of the application of pangenome approaches at different levels of phylogenetic resolution

Level	Organism	Approach[a]	# Genomes	Core size (# genes)	Year (reference)
Species	*Streptococcus agalactiae*	ORFsim, Comb	8	1806	Tettelin et al. (2005)
	Neisseria meningitidis	ORFsim, Comb	6	1337	Schoen et al. (2008)
		ORFsim, Comb	20	1630	Budroni et al. (2011)
	Borrelia burgdoferi	ORFsim, Comb	21	1200	Mongodin et al. (2013)
	Escherichia coli	ORFsim, Comb	17	2344	Rasko et al. (2008)
	Enterococcus faecium	ORFsim, Comb	7	2172	van Schaik et al. (2010)
	Yersinia pestis	ORFsim, Comb	14	3668	Eppinger et al. (2010)
	Streptococcus pyogenes	OG, Comb	11	1376	Lefebure and Stanhope (2007)
	Clostridium difficile	OG, Comb	15	1033	Scaria et al. (2010)
	Lactobacillus paracasei	OG	34	1800	Smokvina et al. (2013)
	Campylobacter jejuni	ORFsim, Ref	130	1042	Meric et al. (2014)
	Campylobacter coli	ORFsim, Ref	62	947	Meric et al. (2014)
	Haemophilus influenzae	FSM	13	1450	Hogg et al. (2007)
	Streptococcus pneumoniae	FSM	17	1400	Hiller et al. (2007)
		ORFsim, Comb	44	1666	Donati et al. (2010)
	Staphylococcus aureus	FSM	16	2245	Boissy et al. (2011)
	Moraxella catarrhalis	FSM	12	1755	Davie et al. (2011)
	Lactobacillus casei	FSM	17	1715	Broadbent et al. (2012)
	Gardnerella vaginalis	FSM	17	746	Ahmed et al. (2012)
	Clostridium botulinum	ORFsim, Comb	13	2657	Bhardwaj and Somvanshi (2017)
Group	*Bacillus cereus*	ORFsim, Comb	4	3000	Lapidus et al. (2008)
	Bacillus subset of species	ORFsim, Comb	12	2009	Eppinger et al. (2011)

(continued)

Table 1 (continued)

Level	Organism	Approach[a]	# Genomes	Core size (# genes)	Year (reference)
Genus	*Streptococcus*	OG, Comb	26	600	Lefebure and Stanhope (2007)
		ORFsim, Comb	52	522	Donati et al. (2010)
	Prochlorococcus	ORFsim, Comb	12	1273	Kettler et al. (2007)
	Bifidobacterium	ORFsim, Comb	14	967	Bottacini et al. (2010)
	Listeria	BMM	13	2032	den Bakker et al. (2010)
	Salmonella	BMM	35	2811	Jacobsen et al. (2011)
	Shewanella	OG	24	1878	Zhong et al. (2018)
	Finegoldia	OG	12	1202	Brüggemann et al. (2018)
Class	Bacilli	IMGM	172	143	Collins and Higgs (2012)
Phylum	Chlamydiae	OG	19	560	Collingro et al. (2011)
Super kingdom	Eubacteria	Gene freq.	573	250	Lapierre and Gogarten (2009)

[a] *ORFsim* ORF alignment similarity, *Comb* combinatorial approach of adding successive genomes, *OG* ortholog clusters, *Ref* initial generation of a reference pangenome using a subset of strains, *FSM* finite supragenome model, *BMM* binomial mixture model, *IMGM* infinitely many genes model, *Gene freq* gene presence/absence frequency

distribution of forecasting the evolution of the pangenome and core-genome size. Another limiting factor, as the number of genomes becomes higher and higher, is the scalability of all possible genome addition permutations, since the total number of comparisons needed is described from the following function:

$$C = \frac{N!}{(n-1)! \cdot (N-n)!}$$

where C is the total number of comparisons, and N is the total number of genomes.

A workaround to an exhaustive approach is a method of subsampling (Vernikos et al. 2015) the total number of comparisons needed; comparisons are randomly selected making sure that each genome undergoes the same number of comparisons; the trick here is to set the number of possible comparisons to a number that will optimally balance the existing computational power and the target dataset size. Indeed, observations from limited in size datasets, showed that even extreme sampling is still able to model reliably the pangenome bypassing the need to follow an exhaustive all-against-all comparison (Fig. 1) (Vernikos et al. 2015). Additional optimizations can be achieved by exploiting alternative (to the original exponential

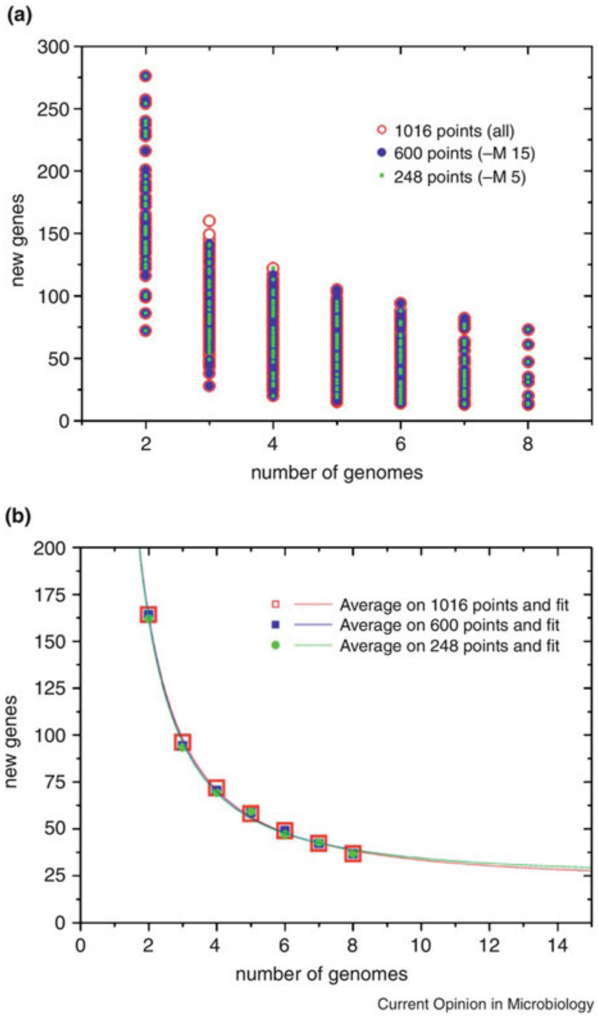


Fig. 1 Pangenome analysis plots for *Streptococcus agalactiae* genomes ($n = 8$). (**a**) Number of new genes detected for adding a genome g to $g - 1$ genomes. Red bubbles: 1016 points for the total number of comparisons (no subsampling). Blue bubbles: 600 points (subsampling, multiplicity of 15). Green bubbles: 248 points (subsampling, multiplicity of 5). (**b**) Regression curve on averages (the subsampling method has limited impact on the outcome)

decay) regressions functions; practical implementations of such optimizations are described in Tettelin et al. (2008), Eppinger et al. (2010, 2011), Mongodin et al. (2013) and Riley et al. (2012).

Recently several stand-alone or server-based suites have become available for pangenome analysis; in the next paragraphs, we will review the most promising and interesting initiatives. See also Table 2 for additional details.

Table 2 Pangenome software synopsis

	Roary	GET_HOMOLOGUES	EDGAR	ITEP	Harvest	PanOCT	panX	PGAP	PanGP	PanCGHweb	SplitMEM
Software type	Package/ Module	Package/Module		Framework/ Library	Toolkit/Suite	Package/ Module		Pipeline/ Workflow			Package/ Module
Interface	Command line interface	Command line interface	Web user interface	Command line interface	Command line interface	Command line interface	Web user interface	Command line interface	Graphical user interface	Web user interface	Command line interface
Input data	Annotated assemblies										
Input format	GFF3				GGR, FASTA, VCF						
Operating system	Unix/ Linux	Unix/Linux, Mac OS		Unix/Linux	Unix/Linux, Mac OS	Unix/ Linux		Unix/Linux	Windows/ Linux		Unix/Linux
Programming languages	Perl	Perl, R	Javascript, R	Python, Shell (Bash)	C++, Python, Shell (Bash)	Perl		Perl	C++		C++
License	GNU General Public License version 2.0	GNU General Public License		GNU General Public License version 2.0		GNU General Public License version 2.0					Apache License version 2.0
Computer skills	Advanced	Advanced	Basic	Advanced	Advanced	Advanced	Basic	Advanced	Basic	Basic	Advanced
	PanViz	EUPAN	PanTools	Spine	AGEnt	Panseq	PanWeb	PanGet	micropan	Pan-Tetris	ClustAGE
Software type	Package/ Module	Toolkit/Suite	Package/ Module	Package/Module	Package/ Module	Package/ Module		Application/ Script	Package/ Module	Framework/ Library	
Interface	Command line interface	Command line interface	Graphical user interface	Command line interface	Command line interface	Command line interface	Web user interface	Command line interface	Command line interface	Graphical user interface	Web user interface and Command line interface
Input data							An annotation files for each genome				
Input format							EMBL				
Operating system	Unix/ Linux	Unix/Linux	Unix/ Linux, Mac OS, Windows	Unix/Linux, Mac OS, Windows	Unix/Linux, Mac OS, Windows	Unix/ Linux, Mac OS, Windows		Unix/Linux, Mac OS, Windows	Unix/ Linux, Mac OS, Windows	On any machine with a Java VM installed	

Programming languages	Javascript	C++, Perl, R	Java	Perl	Perl	Perl	PHP, R	Perl	R	Java	Perl
License	GNU General Public License version 2.0			GNU General Public License version 2.0	GNU General Public License version 2.0	GNU General Public License version 3.0			GNU General Public License version 2.0		GNU General Public License version 3.0
Computer skills	Advanced	Advanced	Medium	Advanced	Advanced	Advanced	Basic	Advanced	Advanced	Medium	Basic
	DeNoGAP	PanACEA	LS-BSR	PANNOTATOR	BGDMdocker	PanCake	Panaconda	PanFunPro	Piggy	seq-seq-pan	NGSPanPipe
Software type	Pipeline/ Workflow	Toolkit/Suite	Pipeline/ Workflow		Pipeline/ Workflow	Package/ Module	Toolkit/ Suite	Application/ Script	Pipeline/ Workflow	Application/ Script	Script
Interface	Command line interface	Command line interface	Command line interface	Web user interface	Command line interface	Command line interface	Command line interface	Web user interface and command line	Command line interface	Command line interface	Command line interface
Input data							An annotation of genome features with some family designation	A list of genomes	Bacterial genome assemblies		Reads in FASTQ format, reference sequence file (FASTA format), parameter for filtering of reads and reference genome protein translation table (PTT) file
Input format								FASTA	GFF3		
Operating system	Unix/ Linux	Unix/Linux, Mac OS, Windows	Unix/ Linux		Unix/Linux, Mac OS, Windows	Unix/ Linux, Windows	Unix/ Linux		Unix/ Linux	Unix/Linux	
Programming languages	Perl	Perl				Python	Python	Perl	Perl, R, Shell (Bash)	Java, Python	Perl
License	GNU General Public License version 3.0	GNU General Public License version 2.0			GNU General Public License version 3.0				GNU General Public License version 3.0	FreeBSD	
Computer skills	Advanced	Advanced	Advanced	Basic	Advanced	Advanced	Advanced	Basic	Advanced	Advanced	Medium

3 Bayesian Decision Model

van Tonder et al. (2014) designed a methodology based on Bayesian decision model, able to analyze directly next-generation sequencing (NGS) data. The model defines the core-genome of bacterial populations allowing also the identification of novel genes. A nice caveat of this approach is that it can analyze even strains without a subset of genes since the model does not assume that all sequences have the entire core gene dataset present. The model has been benchmarked analyzing *Streptococcus pneumoniae* sequences.

4 BGDMdocker

BGDMdocker (Cheng et al. 2017) relies on docker technology to analyze and visualize bacterial pangenome and biosynthetic gene clusters. The pipeline consists of three stand-alone tools, namely Prokka v1.11 (Seemann 2014) for rapid prokaryotic genome annotation, panX (Ding et al. 2018) for pangenome analysis, and antiSMASH3.0 (Weber et al. 2015) for automatic genomic identification and analysis of biosynthetic gene clusters. The visualization supports several options, including alignment, phylogenetic trees, mutations mapped on the phylogenetic branches, and gene loss and gain mapping on the core-genome phylogeny. Benchmarking took place on 44 *Bacillus amyloliquefaciens* strains.

5 Bacterial PanGenome Analysis

Bacterial Pangenome Analysis (BPGA) (Chaudhari et al. 2016), comes with a handful of new options and features most notably that of optimizing the speed of execution. In addition, it offers various entity (core-, pangenome, and MLST) phylogeny, phyletic profile analysis (gene presence/absence), subset analysis, atypical sequence composition analysis, orthologous, and functional annotation for all gene datasets, user-selection of gene clustering algorithm, command line interface, and nice graphics. It runs both in Windows and in Linux as executables files (source code in Perl). BPGA has dependencies with other tools that require installation. In terms of input files, BPGA can “digest” the following file formats: GenBank (.gbk) files, protein sequence file (e.g.,.faa or .fsa or fasta format), binary (0,1) matrix (tab-delimited) file as output of other tools. The seven functional modules of BPGA algorithm include: Pangenome profile analysis, pangenome sequence extraction, exclusive gene family analysis, atypical GC content analysis, pangenome functional analysis, species phylogenetic analysis, and subset analysis.

6 ClustAGE

ClustAGE (Ozer 2018) suite (both online and stand-alone) clusters noncore accessory sequences within a collection of bacterial isolates implementing the BLAST algorithm. It is therefore focused on the accessory genomic dimension of pangenome; Benchmarking of this tool has taken place on *Pseudomonas aeruginosa* genome sequences.

7 DeNoGAP

DeNoGAP (Thakur and Guttman 2016) does many more than pure pangenome analysis, including functional annotation, gene prediction, protein classification, and orthology search; therefore, it is applicable both for complete and draft genomic data. To do this, it implements a big set of existing analysis algorithms. In terms of scalability, it runs linearly due to implementation of iteratively refined Hidden Markov models. Its modular structure supports easy updates and addition of new tools.

8 EDGAR

Implementing phylogenetic concepts like average amino acid and nucleotide identity indices, an online application namely "EDGAR" (Blom et al. 2009, 2016) was developed to support comparative genomic analyses of related isolates. Strong utilities of the suite include Venn diagrams and interactive synteny plots, as well as ease of access to taxa of interest and quick analyses like pangenome vs. core plot, the core-genome and singletons.

9 EUPAN

EUPAN (Hu et al. 2017) is one of the first concrete attempts to analyze eukaryotic pangenomes, even at a relatively low sequencing depth supporting gene annotation of pangenomic dataset, genome assembly, and identification of core and accessory gene datasets exploiting read coverage. The tool has been benchmarked using 453 rice genomes.

10 GET_HOMOLOGUES

GET_HOMOLOGUES (Contreras-Moreira and Vinuesa 2013) is a customizable and detailed pangenome analysis platform (open source written in Perl and R) for microorganisms addressed to non-bioinformaticians. GET_HOMOLOGUES can cluster homologous gene families using bidirectional best-hit clustering algorithms. The cluster granularity can be adjusted by the user based on various filtering strategies (e.g., by controlling key blast parameters such as percentage overlap and identity of pairwise alignments and E-score cutoff value). To estimate the size of the core- and pangenome, the tool supports both exponential and binomial mixture models to fit the data.

11 Harvest

Harvest (Treangen et al. 2014) is suitable for the analysis of (up to thousands of) microbial genomes. It hosts three modules, namely *Parsnp* (core-genome analysis), *Gingr* (output visualization), and *HarvestTools* (meta-analysis). Parsnp exploits jointly whole-genome alignment and read mapping to optimize accuracy and scalability aspects of sequence alignment; this approach can accommodate scalability for up to thousands of genomic datasets. For indexing purposes, it implements directed acyclic graph improving the identification of unique matches (anchors). The input of Parsnp is a directory of MultiFASTA files; the output includes core-genome alignment, variant calls, and a SNP tree, all of which can be visualized via Gingr. Broadly speaking, this tool represents a compromise between whole-genome alignment and read mapping. Parsnp performance has been evaluated on simulated and real data.

12 ITEP

ITEP (Benedict et al. 2014) is a suite of BASH scripts and Python libraries that interface with an SQLite database backend and a large number of tools for the comparison of microbial genomes. ITEP hosts several de novo prediction tools such as sequence alignment, metabolic, clustering, and protein prediction. Users can develop their own customized comparative analysis workflows.

13 LS-BSR

LS-BSR (large-scale BLAST score ratio) (Sahl et al. 2014), calculates a score ratio (BSR value = query/reference bit score) per coding sequence (matrix) within a pangenome dataset using BLAST (Altschul et al. 1997) or BLAT (Kent 2002) for all-against-all alignment purposes. The output (bit score per CDS) can be visualized as a heatmap. Benchmarking has taken place on *Escherichia coli* and *Shigella* datasets.

14 micropan

micropan (Snipen and Liland 2015) is an R package for the pangenome study of prokaryotes. The R computing environment supports several options of statistical analyses (e.g., principal component analysis), pangenome models (e.g., Heaps' law), and graphics. External free software (e.g., HMMER3) is used for the heavy computations involved. Benchmarking has been carried out on 342 *Enterococcus faecalis* genomes.

15 NGSPanPipe

NGSPanPipe (Kulsum et al. 2018) supports microbial pangenome analysis directly from experimental reads. Benchmarking has been carried out using simulated reads of *Mycobacterium tuberculosis*. The pipeline expects as input experimental reads and outputs three files, one of which is a binary matrix showing the presence/absence of genes in each strain; this matrix can be used as input to other pangenome tools like PanOCT (Fouts et al. 2012) and PGAP (Zhao et al. 2012).

16 PanACEA

PanACEA (Clarke et al. 2018) is an open source stand-alone computer program written in Perl that supports users to create an interconnected set of html, javascript, and json files visualizing prokaryotic pan-chromosomes (core and variable regions) generated by PanOCT (Fouts et al. 2012) or other pangenome clustering tools. PanACEA was developed to serve as an intuitive, easy-to-use, stand-alone viewer. Regions and genes can be functionally annotated to allow for visual identification of regions of interest. PanACEA's memory and time requirements are within the capacities of standard laptops. Benchmarking took place on 219 *Enterobacter hormaechei* genomes.

17 Panaconda

Panaconda (Warren et al. 2017) creates whole-genome multiple sequence comparisons and provides a model for representing the relationship among sequences as a graph of syntenic gene families, by discovering collision points within a group of genomes. The first step is to create a de Bruijn graph and use its traversal to build a pan-synteny graph; the alphabet used is based on gene families (instead of nucleotide alphabet). This approach is novel in the context of generating a graph, wherein all sequences are fully represented as paths.

18 PanCake

PanCake (Ernst and Rahmann 2013) is another tool for pangenome analysis (core and unique regions) relying exclusively on sequence data and pairwise alignments (nucmer or BLAST), which makes it annotation independent (i.e., it processes pure whole-genome content). It hosts a command line interface with several subcommands, allowing to add chromosomes, to specify a genome for each chromosome, to add alignments, to compute core and unique regions, and to output selected regions of the analyzed chromosomes. Benchmarking took place on three genera, namely *Pseudomonas*, *Yersinia*, and *Burkholderia*. PanCake is written in Python.

19 PanFunPro

PanFunPro (Lukjancenko et al. 2013) exploits functional information (profiles) for pangenome analysis. The suite supports among others calculation of core, and accessory gene datasets, homology search (all-against-all and pairwise sub-querying), functional annotation (HMM-based), and gene-ontology information analysis. PanFunPro is available both as a standalone (Perl) tool and as a web server. Benchmarking took place on 21 *Lactobacillus* genomes.

20 PanGeT

PanGeT (Yuvaraj et al. 2017) can digest both genomic and proteomic data in order to construct the pangenome for a selection of taxa, exploiting BLASTN or BLASTP, respectively. In terms of performance, it has been benchmarked using a set of 11 *Streptococcus pyogenes* strains. The output is given in the form of a flower plot (core, dispensable, and strain-specific genes).

21 PanGFR-HM

PanGFR-HM (Chaudhari et al. 2018), is putting an interesting view point on the "table" of pangenome, by analyzing exclusively microbes from the Human Microbiome Project; it is a web-based platform integrating functional and genomic analysis for a collection of ~1300 complete human-associated microbial genomes exploiting a novel dimensionality of analysis that of body site (location of the bug in the human body) when comparing different groups of organisms.

22 PanGP

PanGP (Zhao et al. 2014) supports scalable pangenome analysis by analyzing clusters of orthologs pre-computed by OrthoMCL (Li et al. 2003), PGAP (Zhao et al. 2012), Mugsy-Annotator (Angiuoli et al. 2011), or PanOCT (Fouts et al. 2012). In order to predict core and accessory gene datasets, the suite implements random or distance-guided sampling; in the latter, the genomic diversity (GD) drives the sampling of strain permutations. GD is modeled relying on three alternative assumptions: GD is determined by the evolutionary distance on phylogenetic trees, the difference in gene numbers per strain, or by the discrepancy among gene clusters; among the three models the third seems more reliable (preferred model for PanGP).

23 PANINI

PANINI (Abudahab et al. 2018) is a web browser implementation for rapid online visualization and analysis of the core and accessory genome content, implementing unsupervised machine learning with stochastic neighbour embedding based on the t-SNE (t-distributed stochastic neighbour embedding) algorithm; this algorithm calculates first the similarities between the data (in high dimensional space) and then it minimizes the divergence between the two probability matrices over the embedding coordinates. PANINI expects as input the output of Roary (Page et al. 2015).

24 PANNOTATOR

PANNOTATOR (Santos et al. 2013) supports the efforts of automatic annotation transfer onto related unannotated genomes exploiting the existing annotation of a curated genome. From this perspective, it is not a main pangenome analysis tool, but rather as a side-product of cross-comparison it provides pangenomic-related information. Its main contribution though to pangenome analysis is to accelerate the functional annotation of closely related isolates. For this task, it implements a relational database, interactive tools, several SQL reports, and a web-based interface. The expected input is the DNA strand, the gene prediction plus the reference annotated genome.

25 PanOCT

PanOCT (Fouts et al. 2012) is a graph-based ortholog clustering tool for pangenome analysis of closely related prokaryotic genomes exploiting conserved gene neighborhood information to separate recently diverged paralogs into distinct clusters of orthologs where homology-only clustering methods cannot. PanOCT is utilizing BLAST (Altschul et al. 1997) and conserved gene neighborhood information. Four input files are expected including a tabular file of all-versus-all BLASTP searches and the actual protein fasta sequences. PanOCT is specifically designed for pangenome analysis of closely related taxa (in order to be able to distinguish groups of paralogs into separate clusters of orthologs). In terms of memory requirements, PanOCT is greedier than other tools used to benchmark its performance; the memory usage is unchanged until the sixth genome, with a usage of 0.25 GB per genome, maxing out at 0.5 GB per genome by the 25th genome.

26 Panseq

Panseq (Laing et al. 2010) builds pangenomes and identifies single nucleotide polymorphisms (SNPs) using genomic data as input. In addition, it produces files for further phylogenetic analysis exploiting both the information of SNPs as well as the phyletic profile of accessory sequences; all these wrapped-up with a user-friendly graphical user interface.

27 Pan-Tetris

Pan-Tetris (Hennig et al. 2015) is a Java-based tool that exploits an aggregation technique inspired by the Tetris game, to provide an interactive and dynamic visualization of the gene content in a pangenome table with the option of editing and on-the-fly modification of user-defined (pan) gene groups. The suite has been tested on 32 *Staphylococcus aureus* genomes. Pan-Tetris is one of the first attempts that enable modification of the computed pangenome. The computation of whole genome alignment exploits progressiveMAUVE (Darling et al. 2010) algorithm.

28 PanTools

PanTools (Sheikhizadeh et al. 2016) suite supports the construction and visualization of pangenomes hosting online tools and algorithms; the visual representation of the pangenome is based on generalized De Bruijn graphs. The pangenome construction algorithm scales nicely even with large eukaryotic datasets. In addition to the basic pangenome tasks (construction and visualization), the suite supports other handy utilities such as adding, retrieving and grouping of sequences as well as annotating, reconstructing, and comparing genomes or pangenomes. Overall, it can easily support multi-genome read mapping, pangenome browsing, structure-based variation detection and comparative genomics. It has been benchmarked on *E. coli*, yeast, and *Arabidopsis thaliana* genomes.

29 PanViz

PanViz (Pedersen et al. 2017) is a pangenome visualization tool with some analysis options. It can generate dynamic visualizations supporting both pangenome subset selection as well as mapping of new genomes to existing pangenomes. The input data needed is a pangenome matrix (gene group presence/absence across the included genomes), as well as a gene ontology-based functional annotation of each gene group.

30 PanWeb

PanWeb (Pantoja et al. 2017) is a web application that performs pangenome analyses based on PGAP pipeline, providing in addition a user-friendly graphical interface supporting multiple user-defined analysis queries. It can be implemented by users without computational skills. As input, it receives the annotation files for each genome in EMBL format. A complete set of graphs (e.g., pangenome, accessory, core-genome, and unique genes) is provided.

31 panX

panX (Ding et al. 2018) identifies orthologous gene clusters in pangenomes via a user-friendly and interactive web-based visualization. The visualization consists of connected components that allow further analysis. The suite provides alignment and, phylogenetic tree, it maps mutations of each gene cluster and infers gene gain and loss in the core-genome phylogeny. The pipeline breaks annotated genomes into genes and then clusters them into orthologous groups. To identify homologous proteins, panX performs an all-against-all similarity search, while the actual clustering of orthologous genes is carried out by a Markov clustering algorithm.

32 PGAdb-Builder

PGAdb-builder (Liu et al. 2016), constructs a pangenome allele database (PGAdb) to empower whole genome multilocus sequence typing (wgMLST) analyses and operates as a web service suite. Two modules are implemented, namely *Build_PGAdb* for building a PGAdb database and *Build_wgMLSTtree* for constructing a wgMLST tree and determine the genetic relatedness of the input sequences; both modules "digest" genome contigs in FASTA format. PGAdb-builder, has however dependencies with other existing suites like Prokka (Seemann 2014) and Roary (Page et al. 2015).

33 PGAP

PGAP (Zhao et al. 2012) supports pangenome analysis and in addition analysis of functional gene clusters, species evolution, genetic variation, and functional enrichment of query sequences. It outputs the basic pangenome structure and growth curve and in addition SNP and genomic variation information, phylogenetic, and functional annotation metadata. Benchmarking has taken place on *Streptococcus pyogenes* datasets.

34 PGAP-X

Building on PGAP, and in order to more effectively interpret and visualize the results, PGAP-X (Zhao et al. 2018) was developed. The visualization utility can intuitively lead to conclusions on pangenomic structure, conserved regions and overall on genetic variability throughout the pangenomic datasets at hand. Benchmarking has taken place on *S. pneumoniae* and *Chlamydia trachomatis* datasets. One current limitation of PGAP-X (that is not present in PGAP) is that it expects as input only complete genomes.

35 Piggy

Piggy (Thorpe et al. 2018) is a tool for analyzing the intergenic component of bacterial genomes and it is designed to be used in conjunction with Roary (Page et al. 2015). The latter works by analyzing protein-coding sequences thus excluding nonprotein-coding intergenic regions (IGRs) which typically account for approximately 15% of the genome. Piggy matches Roary except that it is based only on IGRs. Benchmarking took place on *Staphylococcus aureus* and *Escherichia coli* using large genome datasets. In terms of input and output, Piggy uses the same format as in Roary and has similar running time requirements. Piggy provides a means to rapidly identify IGR switches, with many evolutionary applications including analysis of the role of horizontal transfer in shaping the bacterial regulome.

36 pyseer

pyseer (Lees et al. 2018), is geared toward genome-wide association studies in the "world" of microbes with the task at hand to identify potential genetic variation linked with certain phenotypic aspects. Pyseer is actually a python implementation of a previous initiative written in C++, namely SEER (Lees et al. 2016). The foundation of pyseer is the use of K-mers (words) of variable length (input) coming from draft assemblies, while using a generalized linear model for each word their link with a potential phenotype is evaluated. In addition, multidimensional scaling of a pairwise distance matrix is implemented in order to control for population structure (embedded in the regression analysis).

37 Roary

Roary (Page et al. 2015) enables the construction of large pangenomes even on a typical desktop machine, yielding fairly accurate output. For example, it can digest up to 1000 strains (13 GB of RAM) building the pangenome in ~4 h. Roary achieves high accuracy which is attributable to utilization of the context of conserved gene neighborhood information. A suite of command line tools is provided to interrogate the dataset providing union, intersection, and complement.

38 seq-seq-pan

seq-seq-pan (Jandrasits et al. 2018) is a workflow for the sequential alignment of sequences to build a pangenome data structure and a whole-genome alignment. seq-seq-pan builds a pangenome data structure allowing editing (addition or removal) of genomes from a set of aligned sequences and subsequent re-alignment of the whole-genome sequences; for whole-genome alignments it relies on progressiveMauve (Darling et al. 2010). The alignment is optimized for generating a representative linear presentation of the aligned set of genomes.

39 Spine and AGEnt

Spine (Ozer et al. 2014) determines the core-genome from a group of genomic sequences and AGEnt (Ozer et al. 2014) identifies the accessory genome in draft genomic sequences. They both use nucmer to align sequences. The pipeline has been tested on genome sequences of *Pseudomonas aeruginosa*. However, as mentioned by the authors, whole genome alignment of reference genomes and core-genome identification with Spine can be time-consuming.

40 SplitMEM

SplitMEM (Marcus et al. 2014) scales linearly in terms of time and space in relation to the number of genomes of interest. To do this, it traverses suffix trees (for the genomes) and builds compressed de Bruijn graphs of pangenomes. In terms of notation, nodes within the graph represent conserved or strain-specific sequences of the pangenome. Benchmarking has taken place on *Bacillus anthracis* and *E. coli* datasets.

41 Highlights

Pangenome analysis has today many options when it comes to practical implementation. Depending on the analysis focus, the desired input and output, the dependability on other algorithms, as well as the modeling parametrization, users have many options to choose from. In the current review, we highlight the following five tools: BPGA (Chaudhari et al. 2016) for its very fast execution time, the intuitive handling and the user-defined clustering algorithm, Roary (Page et al. 2015) due to its internal processing (clustering of high similarity sequences) that results in linear memory consumption, LS-BSR (Sahl et al. 2014) that similarly to Roary performs pre-clustering reducing substantially the running time, PanOCT (Fouts et al. 2012), which takes into account both homology and positional gene neighborhood information and PGAP (Zhao et al. 2012) that can work also with draft forms of genomic data such as annotated assemblies.

42 Food for Thought

The final results and conclusions of a pangenome analysis, among others, massively depend on the following aspects, that need thoughtful consideration prior to embarking any such project: Homology search algorithm, the phylogenetic sample at hand, the pangenome model implemented and the type and quality of sequence entities (e.g., DNA, protein, presence/absence—phyletic profile, and SNPs).

For example, when it comes to homology definition based on sequence similarity there is a wide range of similarity thresholds used in previous attempts: $i = 50\%$, $L = 50\%$ (Tettelin et al. 2005), $i = 70\%$, $L = 70\%$ (Hiller et al. 2007), $i = 70\%$, $L = 50\%$ (Meric et al. 2014), $i = 30\%$, $L = 80\%$ (Bentley et al. 2007), where i stands for sequence identity and L for sequence length.

The starting level (ORFs, CDSs, genes, proteins, SNPs) and the quality (in silico, manual curation) of annotation as well as inherent bacterial genomic complexity at the sequence level such as low complexity repeats, recombination hot spots, horizontally acquired genomic fragments constitute other important aspects of consideration. Such information variability can massively affect the predicted conserved and unique genes in favor of the former or the latter; this might also determine the structure of pangenome (open or closed).

43 Conclusions

Being able algorithmically to digest the largest possible pool of data available is critical in order to approach more reliably the phylogenetic history of bacterial populations. Indeed such comparative genomic analyses started by exploiting ~0.07% of a genome (16s rRNA) (Woese 1987), latter on using up to ~0.2% of the genomic information (MLST) (Maiden et al. 1998), and recently up to 100% of the information exploiting the pangenome wealth of data (Medini et al. 2005; Tettelin et al. 2005).

The recent explosion of sequencing projects replaced the limiting factor of *data sparsity* with the *immense data dimensionality* (Vernikos 2010) and we are now in the middle of a transformation moving from top-down (trying to fit the limited data to the model) to bottom-up approaches in an attempt to move from the "infant" stage of single-strain genomics to the post pangenome era of "adulthood." The model assumptions therefore become less and less pivotal as the pace of primary data generation continues to grow exponentially, asking not for modeling superpower but instead interpretation and connecting the dots super skills.

References

Abudahab K, Prada JM, Yang Z, Bentley SD, Croucher NJ, Corander J, Aanensen DM (2018) PANINI: pangenome neighbour identification for bacterial populations. Microb Genom 5(4). https://doi.org/10.1099/mgen.0.000220

Ahmed A, Earl J, Retchless A, Hillier SL, Rabe LK, Cherpes TL, Powell E, Janto B, Eutsey R, Hiller NL et al (2012) Comparative genomic analyses of 17 clinical isolates of Gardnerella vaginalis provide evidence of multiple genetically isolated clades consistent with subspeciation into genovars. J Bacteriol 194(15):3922–3937

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Angiuoli SV, Dunning Hotopp JC, Salzberg SL, Tettelin H (2011) Improving pan-genome annotation using whole genome multiple alignment. BMC Bioinf 12:272

Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, Price ND (2014) ITEP: an integrated toolkit for exploration of microbial pan-genomes. BMC Genomics 15:8

Bentley SD, Vernikos GS, Snyder LA, Churcher C, Arrowsmith C, Chillingworth T, Cronin A, Davis PH, Holroyd NE, Jagels K, Maddison M, Moule S, Rabbinowitsch E, Sharp S, Unwin L, Whitehead S, Quail MA, Achtman M, Barrell B, Saunders NJ, Parkhill J (2007) Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. PLoS Genet 3(2):e23

Bhardwaj T, Somvanshi P (2017) Pan-genome analysis of Clostridium botulinum reveals unique targets for drug development. Gene 623:48–62. https://doi.org/10.1016/j.gene.2017.04.019

Blom J, Albaum SP, Doppmeier D, Puhler A, Vorholter FJ, Zakrzewski M, Goesmann A (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinf 10:154

Blom J, Kreis J, Spanig S, Juhre T, Bertelli C, Ernst C, Goesmann A (2016) EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Res 44(W1): W22–W28

Boissy R, Ahmed A, Janto B, Earl J, Hall BG, Hogg JS, Pusch GD, Hiller LN, Powell E, Hayes J et al (2011) Comparative supragenomic analyses among the pathogens Staphylococcus aureus, Streptococcus pneumoniae, and Haemophilus influenzae using a modification of the finite supragenome model. BMC Genomics 12:187

Bottacini F, Medini D, Pavesi A, Turroni F, Foroni E, Riley D, Giubellini V, Tettelin H, van Sinderen D, Ventura M (2010) Comparative genomics of the genus Bifidobacterium. Microbiology 156(Pt 11):3243–3254

Broadbent JR, Neeno-Eckwall EC, Stahl B, Tandee K, Cai H, Morovic W, Horvath P, Heidenreich J, Perna NT, Barrangou R et al (2012) Analysis of the Lactobacillus casei supragenome and its influence in species evolution and lifestyle adaptation. BMC Genomics 13:533

Brüggemann H, Jensen A, Nazipi S, Aslan H, Meyer RL, Poehlein A, Brzuszkiewicz E, Al-Zeer MA, Brinkmann V, Söderquist B (2018) Pan-genome analysis of the genus Finegoldia identifies two distinct clades, strain-specific heterogeneity, and putative virulence factors. Sci Rep 8 (1):266. https://doi.org/10.1038/s41598-017-18661-8

Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV et al (2011) Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci U S A 108(11):4494–4499

Chaudhari NM, Gupta VK, Dutta C (2016) BPGA- an ultra-fast pan-genome analysis pipeline. Sci Rep 6:24373

Chaudhari NM, Gautam A, Gupta VK, Kaur G, Dutta C, Paul S (2018) PanGFR-HM: a dynamic web resource for pan-genomic and functional profiling of human microbiome with comparative features. Front Microbiol 9:2322

Cheng G, Quan L, Zhou Z, Ma L, Zhang G, Wu Y, Chen C (2017) BGDMdocker: an workflow base on Docker for analysis and visualization pan-genome and biosynthetic gene clusters of bacterial. bioRxiv:098392

Clarke TH, Brinkac LM, Inman JM, Sutton G, Fouts DE (2018) PanACEA: a bioinformatics tool for the exploration and visualization of bacterial pan-chromosomes. BMC Bioinf 19(1):246

Collingro A, Tischler P, Weinmaier T, Penz T, Heinz E, Brunham RC, Read TD, Bavoil PM, Sachse K, Kahane S et al (2011) Unity in variety — the pan-genome of the Chlamydiae. Mol Biol Evol 28(12):3253–3270

Collins RE, Higgs PG (2012) Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol Biol Evol 29(11):3413–3425

Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79(24):7696–7701

Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5(6):e11147

Davie JJ, Earl J, de Vries SP, Ahmed A, Hu FZ, Bootsma HJ, Stol K, Hermans PW, Wadowsky RM, Ehrlich GD et al (2011) Comparative analysis and supragenome modeling of twelve Moraxella catarrhalis clinical isolates. BMC Genomics 12:70

den Bakker HC, Cummings CA, Ferreira V, Vatta P, Orsi RH, Degoricija L, Barker M, Petrauskene O, Furtado MR, Wiedmann M (2010) Comparative genomics of the bacterial genus Listeria: genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics 11:688

Ding W, Baumdicker F, Neher RA (2018) panX: pan-genome analysis and exploration. Nucleic Acids Res 46(1):e5

Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, Oggioni M, Dunning Hotopp JC, Hu FZ, Riley DR et al (2010) Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. Genome Biol 11(10):R107

Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, Mou S, Achtman M, Lindler LE, Ravel J (2010) Genome sequence of the deep-rooted Yersinia pestis strain Angola reveals new insights into the evolution and pangenome of the plague bacterium. J Bacteriol 192(6):1685–1699

Eppinger M, Bunk B, Johns MA, Edirisinghe JN, Kutumbaka KK, Koenig SS, Creasy HH, Rosovitz MJ, Riley DR, Daugherty S et al (2011) Genome sequences of the biotechnologically important Bacillus megaterium strains QM B1551 and DSM319. J Bacteriol 193(16):4199–4213

Ernst C, Rahmann S (2013) PanCake: a data structure for pangenomes. German Conference on Bioinformatics, Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269(5223):496–512

Fouts DE, Brinkac L, Beck E, Inman J, Sutton G (2012) PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res 40(22):e172

Hennig A, Bernhardt J, Nieselt K (2015) Pan-Tetris: an interactive visualisation for pan-genomes. BMC Bioinf 16(Suppl 11):S3

Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, Keefe R, Ehrlich NE, Shen K, Hayes J et al (2007) Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. J Bacteriol 189(22):8186–8195

Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD (2007) Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol 8(6):R103

Hu Z, Sun C, Lu KC, Chu X, Zhao Y, Lu J, Shi J, Wei C (2017) EUPAN enables pan-genome studies of a large number of eukaryotic genomes. Bioinformatics 33(15):2408–2409

Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW, Friis C (2011) The Salmonella enterica pan-genome. Microb Ecol 62(3):487–504

Jandrasits C, Dabrowski PW, Fuchs S, Renard BY (2018) Seq-seq-pan: building a computational pan-genome data structure on whole genome alignment. BMC Genomics 19(1):47

Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12(4):656–664

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J et al (2007) Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. PLoS Genet 3(12):e231

Kulsum U, Kapil A, Singh H, Kaur P (2018) NGSPanPipe: a pipeline for pan-genome identification in microbial strains from experimental reads. Adv Exp Med Biol 1052:39–49

Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, Thomas JE, Gannon VP (2010) Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinf 11:461

Lapidus A, Goltsman E, Auger S, Galleron N, Segurens B, Dossat C, Land ML, Broussolle V, Brillard J, Guinebretiere MH et al (2008) Extending the Bacillus cereus group genomics to putative food-borne pathogens of different toxicity. Chem Biol Interact 171(2):236–249

Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. Trends Genet 25 (3):107–110

Lees JA, Vehkala M, Valimaki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Davies MR, Steer AC, Tong SY, Honkela A, Parkhill J, Bentley SD, Corander J (2016) Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun 7:12797

Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J (2018) Pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics 34(24):4310–4312

Lefebure T, Stanhope MJ (2007) Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol 8(5):R71

Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9):2178–2189

Liu YY, Chiou CS, Chen CC (2016) PGAdb-builder: a web service tool for creating pan-genome allele database for molecular fine typing. Sci Rep 6:36213

Lukjancenko O, Thomsen M, Voldby Larsen M, Ussery D (2013) PanFunPro: PAN-genome analysis based on FUNctional PROfiles [version 1; referees: 3 approved with reservations]. F1000Res 2:265

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG (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(6):3140–3145

Marcus S, Lee H, Schatz MC (2014) SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics 30(24):3476–3483

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15(6):589–594

Meric G, Yahara K, Mageiros L, Pascoe B, Maiden MC, Jolley KA, Sheppard SK (2014) A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic campylobacter. PLoS One 9(3):e92798

Mongodin EF, Casjens SR, Bruno JF, Xu Y, Drabek EF, Riley DR, Cantarel BL, Pagan PE, Hernandez YA, Vargas LC et al (2013) Inter- and intra-specific pan-genomes of Borrelia burgdorferi sensu lato: genome stability and adaptive radiation. BMC Genomics 14:693

Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, Thomas AD, Ali R, Sharma K, Kyrpides NC, Reddy TB (2017) Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Res 45(D1):D446–D456

Ozer EA (2018) ClustAGE: a tool for clustering and distribution analysis of bacterial accessory genomic elements. BMC Bioinf 19(1):150

Ozer EA, Allen JP, Hauser AR (2014) Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. BMC Genomics 15:737

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31 (22):3691–3693

Pantoja Y, Pinheiro K, Veras A, Araujo F, Lopes de Sousa A, Guimaraes LC, Silva A, Ramos RTJ (2017) PanWeb: a web interface for pan-genomic analysis. PLoS One 12(5):e0178154

Pedersen TL, Nookaew I, Wayne Ussery D, Mansson M (2017) PanViz: interactive visualization of the structure of functionally annotated pangenomes. Bioinformatics 33(7):1081–1082

Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R et al (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol 190(20):6881–6893

Riley DR, Angiuoli SV, Crabtree J, Dunning Hotopp JC, Tettelin H (2012) Using Sybil for interactive comparative genomics of microbes on the web. Bioinformatics 28(2):160–166

Rodriguez-Valera F, Ussery DW (2012) Is the pan-genome also a pan-selectome? F1000Res 1:16

Sahl JW, Caporaso JG, Rasko DA, Keim P (2014) The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ 2:e332

Santos AR, Barbosa E, Fiaux K, Zurita-Turk M, Chaitankar V, Kamapantula B, Abdelzaher A, Ghosh P, Tiwari S, Barve N, Jain N, Barh D, Silva A, Miyoshi A, Azevedo V (2013) PANNOTATOR: an automated tool for annotation of pan-genomes. Genet Mol Res 12 (3):2982–2989

Scaria J, Ponnala L, Janvilisri T, Yan W, Mueller LA, Chang YF (2010) Analysis of ultra low genome conservation in Clostridium difficile. PLoS One 5(12):e15147

Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, Muller T, Goesmann A, Joseph B, Konietzny S, Kurzai O et al (2008) Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis. Proc Natl Acad Sci U S A 105(9):3473–3478

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30 (14):2068–2069

Sheikhizadeh S, Schranz ME, Akdel M, de Ridder D, Smit S (2016) PanTools: representation, storage and exploration of pan-genomic data. Bioinformatics 32(17):i487–i493

Snipen L, Liland KH (2015) Micropan: an R-package for microbial pan-genomics. BMC Bioinf 16:79

Smokvina T, Wels M, Polka J, Chervaux C, Brisse S, Boekhorst J, van Hylckama Vlieg JE, Siezen RJ (2013) Lactobacillus paracasei comparative genomics: towards species pan-genome definition and exploitation of diversity. PLoS One 8(7):e68731

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278(5338):631–637

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS 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 (39):13950–13955

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11(5):472–477

Thakur S, Guttman DS (2016) A De-Novo Genome Analysis Pipeline (DeNoGAP) for large-scale comparative prokaryotic genomics studies. BMC Bioinf 17(1):260

Thorpe HA, Bayliss SC, Sheppard SK, Feil EJ (2018) Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. Gigascience 7(4):1–11

Treangen TJ, Ondov BD, Koren S, Phillippy AM (2014) The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 15 (11):524

van Schaik W, Top J, Riley DR, Boekhorst J, Vrijenhoek JE, Schapendonk CM, Hendrickx AP, Nijman IJ, Bonten MJ, Tettelin H et al (2010) Pyrosequencingbased comparative genome analysis of the nosocomial pathogen Enterococcus faecium and identification of a large transferable pathogenicity island. BMC Genomics 11:239

van Tonder AJ, Mistry S, Bray JE, Hill DM, Cody AJ, Farmer CL, Klugman KP, von Gottberg A, Bentley SD, Parkhill J, Jolley KA, Maiden MC, Brueggemann AB (2014) Defining the estimated core genome of bacterial populations using a Bayesian decision model. PLoS Comput Biol 10(8):e1003788

Vernikos GS (2010) The pyramid of knowledge. Nat Rev Microbiol 8(2):91

Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154

Warren AS, Davis JJ, Wattam AR, Machi D, Setubal JC, Heath L (2017) Panaconda: application of pan-synteny graph models to genome content analysis. bioRxiv:215988

Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Muller R, Wohlleben W, Breitling R, Takano E, Medema MH (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43(W1): W237–W243

Woese CR (1987) Bacterial evolution. Microbiol Rev 51(2):221–271

Yuvaraj I, Sridhar J, Michael D, Sekar K (2017) PanGeT: pan-genomics tool. Gene 600:77–84

Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J (2012) PGAP: pan-genomes analysis pipeline. Bioinformatics 28(3):416–418

Zhao Y, Jia X, Yang J, Ling Y, Zhang Z, Yu J, Wu J, Xiao J (2014) PanGP: a tool for quickly analyzing bacterial pan-genome profile. Bioinformatics 30(9):1297–1299

Zhao Y, Sun C, Zhao D, Zhang Y, You Y, Jia X, Yang J, Wang L, Wang J, Fu H, Kang Y, Chen F, Yu J, Wu J, Xiao J (2018) PGAP-X: extension on pan-genome analysis pipeline. BMC Genomics 19(Suppl 1):36

Zhong C, Han M, Yu S, Yang P, Li H, Ning K (2018) Pan-genome analyses of 24 Shewanella strains re-emphasize the diversification of their functions yet evolutionary dynamics of metal-reducing pathway. Biotechnol Biofuels 11:193. https://doi.org/10.1186/s13068-018-1201-1

Part II
Evolutionary Biology of Pangenomes

Structure and Dynamics of Bacterial Populations: Pangenome Ecology

Taj Azarian, I-Ting Huang, and William P. Hanage

Abstract Prokaryotes demonstrate tremendous variation in gene content, even within individual bacterial clones or lineages. This diversity is made possible by the ability of bacteria to horizontally transfer DNA through a variety of mechanisms, and the extent of such transfer sets them apart from eukaryotes. What has become evident through interrogation of thousands of bacterial genomes is that gene variation is directly related to the ecology of the organism and is driven by continual processes of niche exploration, diversification, and adaptation. Of course, the acquisition of new genes is not necessarily beneficial, resulting in either the removal of that individual through purifying selection or the occurrence of compensatory mutations in the genomic "backbone" (i.e., core genes) that become epistatically linked to the presence accessory genes. There are now numerous examples of relationship between gene variation and niche adaptation. We explore some of those examples here as well as the population genomic footprint left by the dynamics of gene flow, diversification, and adaptation.

Keywords Bacterial population genomics · Pangenome · Accessory genome · Ecology · Adaptation · Bacterial evolution · Recombination · Horizontal gene transfer

T. Azarian
College of Medicine, University of Central Florida, Orlando, FL, USA

I.-T. Huang · W. P. Hanage (✉)
Center for Communicable Disease Dynamics, Department of Epidemiology, T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
e-mail: whanage@hsph.harvard.edu

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_5

1 Introduction

Pangenomes of bacterial species show a tremendous range of diversity in size, content, and fluidity. In comparing the core genome size in relation to the accessory genome, some species possess relatively limited pangenomes while others are expansive. Accessory genomes may be composed of genes belonging to phages, transposons, insertion sequences, and plasmids, as well as genes that have diverged through mutation and recombination to the point where they are considered as a separate homolog. Some of these genomic elements may be relatively stable (e.g., an integrated prophage), while others may be gained and lost within a single bacterial culture (e.g., plasmids). In this chapter, we will discuss the population genomics of pan-, and more specifically, accessory genomes, specifically detailing how accessory genomes vary among and within bacterial species and the implications this variation has for microbial ecology. Throughout this discussion, it is important to not lose sight of what we are referring to with the catch-all phrase "accessory." These are the dynamic elements of the genome, often containing large genomic islands that augment the bacterium's phenotype, which may, as we will outline, be used to glean knowledge of ecology and evolutionary history of a genus, species, or set of lineages. Further, in no way does the term accessory or the misleading synonym "dispensable" suggest non-essential, as some "accessory" genes actually represent divergent variants of an essential gene.

2 Mechanisms of Pangenome Variation

The content and diversity of a bacteria's accessory genome are directly associated with the mode and frequency of horizontal gene transfer (HGT), which in turn is tightly linked to ecology. Modes of HGT include transformation: the uptake and integration of exogenous DNA from the environment, transduction: the introduction of exogenous DNA into the bacterial cell through a viral vector (e.g., bacteriophage), and conjugation: the direct transfer of DNA between two bacterial cells through a pilus, which usually involves plasmids and transposons. Bacteria vary in the degree to which each of these mechanisms occurs within their populations and in their DNA uptake mechanisms. It is also almost certain that other variants of these mechanisms remain to be discovered, as illustrated by recent work describing "lateral transduction" capable of transferring genomic regions of remarkable size (Chen et al. 2018).

Integrative and conjugative elements (ICE) include integrative plasmids and conjugative transposons, which are circularized mobile elements transferred through conjugation. ICE may harbor a number of genes important to virulence, specialized metabolism, and survival, and are the primary means by which antibiotic-resistant genes are transmitted among bacteria. Plasmids may contain anywhere from 5 to 100 or so genes, allowing for a lineage to gain or lose many loci in a single step, especially for those species with high plasmid diversity. Phylum *Proteobacteria*, which includes several pathogenic species from genera Escherichia, Salmonella, Vibrio, Helicobacter, Yersinia, and Legionellales possess some of the most prevalent and diverse plasmids with a wide host range (Shintani et al. 2015). Therefore, unsurprisingly, species among these genera have moderately large pangenome sizes (McInerney et al. 2017).

Naturally, competent (transformable) species are able to uptake DNA directly from the environment resulting in homologous or nonhomologous recombination, the latter frequently associated with gene gain (Croucher et al. 2012). Arguably, the most famous of these species, *Streptococcus pneumoniae*, was made so by its role in the Griffith experiments in 1928, which led to the identification of DNA as the conveyor of genetic information. Through those experiments, Griffith observed that "smooth" (i.e., unencapsulated) avirulent *S. pneumoniae* could become virulent through exposure to heat-killed virulent "rough" (i.e., encapsulated) pneumococci (Griffith 1928). We now know that what he observed was transformation resulting in the acquisition of the capsular polysaccharide (CPS) loci that code genes responsible for the synthesis and polymerization the antigenic serotype capsule. There are over 90 serotypes identified and the CPS loci span 10,337–30,298 bp with at least 26 coding sequences depending on the particular serotype (Bentley et al. 2006). Therefore, this single recombination event resulted in the acquisition of 26 accessory genes. Since then, other species including *Neisseria gonorrhoeae, Campylobacter jejuni, Vibrio cholerae*, and *Haemophilus influenzae* have been found to be naturally competent.

Another method by which transformation may result in differences in gene content is through events that lead to gene diversification, which are frequently observed among several species as recombination "hotspots." The primary effect of these events is antigenic variation in genes linked to host–pathogen interactions. For example, among pneumococci, two virulence factors, pneumococcal surface proteins A and C (pspA and pspC), are known to have 3 and 11 variants, respectively (Hollingshead et al. 2000; Iannelli et al. 2002). These variants are diverse in length, structural organization, and nucleotide variation, the results of frequent recombination events. Most important, they are different in serology, which has significant implications for host immunity (Azarian et al. 2016; Georgieva et al. 2018). Similarly, among gonococci, the *opa* and neighboring *pil* loci are highly mosaic due to recombination of existing alleles (Bilek et al. 2009). The gene product *Opa* is an outer membrane adhesion protein that is important for colonization and invasion of the genital and nasopharyngeal mucosal epithelium. As a note, antigenic variation through recombination leads to an interesting contradiction in terminology. In both of these examples, pspA, pspC, opa, and pil are considered "core" genes in the sense that each member of their respective species possesses a variant. They are by all definitions "essential" to core cell function; yet, through current methods of pangenome analysis that are commonly based on a nucleotide homology level of at least 80%, they are identified operationally as accessory genes. Finally, transduction through temperate bacteriophages may introduce considerable gene variation in both Gram-negative and Gram-positive bacteria (Feng et al. 2008; Waldor and Friedman 2005). While their precise evolutionary impact in most cases remains unclear, it is certain that their pathogenesis plays a significant role in the biology of their host. For example, many phages harbor genes coding for virulence factors including toxins or secreted enzymes (Romero et al. 2009); therefore, prophages (bacteriophages integrated into host bacterial genomes) represent a significant mechanism for variation of virulence among closely related bacteria (Fortier and Sekulovic 2013). In relation to pangenome dynamics, the transmission of bacteriophages can result in significant variation among bacterial populations on short timescales by two mechanisms: through (1) the direct integration of the prophage and (2) the acquisition or evolution of antiphage mechanisms. The later may involve phage-inducible chromosomal islands and CRISPR-Cas systems (Reyes-Robles et al. 2018), which independently represent instances of gene acquisition and a source of pangenome variation. Predator–prey dynamics of bacteriophages and their host has been widely observed with *Siphoviridae* phages and *S. pneumoniae* (Romero et al. 2009), lamba STX-coding phage in Shiga toxin-producing *E. coli*, and ICP (*Myoviridae*) and CTX phages in *Vibrio cholerae* (Seed et al. 2011; Waldor and Friedman 2005), among myriad others. The result is highly variable prophage content even within closely related members of bacterial lineages (Croucher et al. 2014).

3 Population Genomics of Pangenomes

Today, the identification of a bacterial sample's core genome is a common intermediate step among bioinformatics pipelines for preparing whole-genome sequencing data for phylogenetic analysis. Historically, the accessory genome was largely ignored with the exception of the identification of important genes such as those conferring antibiotic resistance or increased virulence. Methodologically, it was difficult to scale accessory genome analysis to large population samples of a species and especially across several species. Then, the discovery that in three diverse *E. coli* isolates, less than 40% of the genes was found in the genomes of all three demonstrated that extensive variation was possible (Welch et al. 2002). A subsequent study of just eight genomes of *Streptococcus agalactiae* (Group B *Streptococcus*) published in 2005 identified 1806 core genes and 439 "dispensable" genes, highlighting that tremendous variation could be observed with even a small sample (Tettelin et al. 2005). This chapter introduced the concept of the pangenome. Now, large-scale analyses of pangenomes continue to reveal significant diversity even over short timescales, providing information about the demographic history and adaptive evolution of bacteria. These studies have shown that pangenome size and diversity vary among species and depend on lifestyle (McInerney et al. 2017; Ochman and Davalos 2006).

McInerney and colleagues recently summarized the range of diversity observed among bacterial species (McInerney et al. 2017). Pangenome sizes ranged from 974 for the obligate intracellular bacteria *Chlamydia trachomatis* to 40,362 for the semiaquatic agricultural *Oryza sativa*. Comparing sizes of accessory genomes in relation to the total number of genes in the pangenome, *O. sativa* had the smallest, just 8% of genes were accessory, while in *Salmonella enterica*, a staggering 83% of its 10,267 genes are found in the accessory genome. Assessing "genomic fluidity" is another method for quantifying pangenome diversity (Kislyuk et al. 2011). Instead of assessing the relationship between core and accessory genome size, genomic fluidity measures the dissimilarity of genomes evaluated at the gene level calculated as the "ratio of unique gene families to the sum of gene families in pairs of genomes averaged over randomly chosen genome pairs from within a group of N genomes." In a comparison of genomic fluidity among seven species known to undergo HGT, *Neisseria meningitidis, Escherichia coli*, and *Streptococcus* spp. ranked highest in genomic fluidity (Kislyuk et al. 2011) (although it should be noted that this metric is expected to be affected by the sample chosen for study).

Within a species, accessory genome diversity increases with core genome divergence and models of homologous recombination and HGT have shown how these processes lead to the formation of population structure (Croucher et al. 2014; Marttinen et al. 2015). Boundaries for HGT across species roughly follow the same trajectories. Species in genera *Streptococcus*, *Neisseria*, and *Campylobacter*, for example, have been shown to engage in HGT more frequently with closely related members (e.g., between *S. pneumoniae* and *S. mitis* and *S. oralis*, and *N. gonorrhoeae* and *N. meningitidis*). Therefore, the size and distribution of accessory genes in a population provide insights into the demographic history of bacterial species as well as delineations of species boundaries.

As we have described, many methods can generate accessory genome diversity. While not wholly analogous to the way nucleotide mutations arise and propagate in a population, the gain and loss of genes nonetheless inform the shared evolutionary history of a population in the same manner. Genomic islands acquired through HGT often become relatively fixed in bacterial lineages (Croucher et al. 2014) with the number of acquired genes increasing with lineage age (Donati et al. 2010). This is especially true for Staphylococcal Cassette Chromosome mec (*SCCmec*) elements in clones of *S. aureus* (International Working Group on the Classification of Staphylococcal Cassette Chromosome Elements (IWG-SCC) 2009), pathogenicity islands among toxigenic and non-toxigenic lineages of *V. cholerae* (Wozniak et al. 2009), and CPS loci in pneumococci (Bentley et al. 2006). These mobile elements, therefore, inform long-scale evolutionary history, while in the short term, prophage variation and the scars of transformation events reflect more recent events. As such, it is possible to recapitulate the core genome phylogeny of a population through phylogenetic reconstruction using a presence–absence alignment of accessory genes, represented by 1's and 0's, respectively (Azarian et al. 2018). In essence, this represents a tight linkage between core genome single nucleotide polymorphisms and the history of gene gain and loss. This may, of course, oversimplify the complex interconnected processes that led to accessory gene variation, but it does provide an easy data structure that may be investigated to understand how bacterial populations change over time.

An interesting approach to assessing temporal changes in bacterial population genomics is to consider the dynamics of the accessory genome. The clearest examples of this are observations of rapid changes in virulence or antibiotic resistance among bacterial lineages, often leading to short-term success of a clone (Croucher et al. 2014). The impact of human interventions, namely vaccines, affects not only the distribution of lineages in a population but also the available pool of accessory genes. For example, if an ICE is strongly associated with a lineage, and that lineage is targeted by vaccine, then the removal of the lineage from the population may ultimately remove the reservoir for that ICE. The impact of vaccine on the pathogen population of *S. pneumoniae* has been extensively studied (Azarian et al. 2018; Croucher et al. 2013). After the introduction of the seven-valent pneumococcal conjugate vaccine (PCV7) in the USA, an analysis of a sample of 616 genomes of pneumococci carried in children in Massachusetts showed the removal of accessory genes associated with the CPS loci of vaccine serotypes (Croucher et al. 2013). In addition, the prevalence of antibiotic-resistant genes associated with two transposons was shifted due to the removal of two vaccine lineages they were associated with and the subsequent emergence of a non-vaccine lineage harboring one of the transposons. A study of pneumococcal population dynamics over 13 years and spanning the introduction of the PCV7 showed that the introduction of vaccines greatly shifted the frequencies of accessory genes in the population (Azarian et al. 2018). Surprisingly, the frequencies of accessory genes then shifted back to pre-vaccine values as the pneumococcal population recovered from the removal of nearly 30% prevalent genotypes targeted by vaccine. This observation was elucidated by recent work by Corander and colleagues who investigated accessory gene frequencies across of 4127 pneumococcal isolates from four distinct geographic areas (Corander et al. 2017). They found that accessory genes had similar frequencies in the four populations despite significant differences in lineage composition and the timing of vaccine use. Through functional analysis of the accessory genes and population dynamic modeling, they proposed that the frequencies of accessory genes are shaped by negative frequency-dependent selection (NFDS) through pathogen–pathogen, host–pathogen, and pathogen–environment interactions. Classically defined, in an NFDS model the fitness of a phenotype depends on its frequency relative to other phenotypes in a population. The same NFDS model has been used to explain the diversity of protein antigens among pneumococci, which we briefly touched upon early in the chapter. In the case of protein antigens, increasing host immunity toward an antigen drives diversification of the gene coding for the protein either through mutation, or most often, recombination. The same dynamic can be observed with prophages and restriction modification systems that defend against infection. Ultimately, these observations point to a central hypothesis for accessory genome variation, that difference in gene content are linked to adaptation and niche specialization, but that in the case of NFDS the niche may be dynamically generated by fluctuating frequencies of loci in the pangenome.

4 The Ecological Significance of Pangenomes

The observation of pangenomes as a common feature of many bacteria begs the question of what has selected them? What are the ecological features that lead to the pervasive association of a core, with a disseminated complement of many additional genes, some shared with other species? While some have clear selective consequences, most are obscure. The extent to which bacteria vary in gene content sets them apart from eukaryotes, and is just one of the reasons we cannot easily transfer population genetic concepts between the superkingdoms of life. One metaphor for bacteria and their varying genome content compares them to modern smartphones (Young 2016) in which the core genome is the operating system, the accessory genome is the apps downloaded to the phone, and the pangenome would be everything in the app store. In the following, we divide up the accessory genes that combine to make up a pangenome into various categories, not by function but by how they are distributed among lineages in the population.

The perspective we take is of the bacterial genome as a transient construct. Loci can be added to it, and selected to become more common or indeed lost from the population, should they no longer be necessary. The pangenome for any sample is the totality of genes currently associated with its contents. This need not be a permanent or even especially long relationship. Consider a locally prominent prophage, which might not be present in the same population if you returned at a later date. Indeed we can imagine that given the many ways bacteria engage in HGT, a sample of sufficient size will contain many loci in a new genetic background that are yet to be lost (analogous to incomplete purifying selection (Rocha et al. 2006). A subset of the pangenome, expected to be rare in any reasonably large sample, is genes that are either infrequently obtained or actively selected against. In general, the extent of gene flow will be regulated by the genetic and ecological similarity of the bacteria and the compatibility of the genetic background to adapt to the acquisition of novel genes (Wiedenbeck and Cohan 2011).

Moving to loci that are present at intermediate frequencies, say between 5% and 95% of isolates, we can distinguish between loci that are restricted to a few lineages, or are widely disseminated but not fixed in any lineage. These suggest different evolutionary scenarios. Dealing with the latter first; a locus that is easy to obtain but hard to hold onto suggests fluctuating selection. We see it more often than the genes in the previous category, because it provides selective benefits. However, these are not consistent benefits or we would expect the gene to rapidly become more common and indeed part of the core. Examples of these include drug-resistant genes in lineages that lack compensatory mutations, and as such only experience a selective benefit in the presence of the drug (Blanquart et al. 2018; Cobey et al. 2017; Lehtinen et al. 2017).

In contrast, loci fixed in a lineage might represent the ecological "address" of those bacteria, a dimension of their niche. However, this need not be the case. Studies of populations of *S. pneumoniae* have shown that the accessory loci in this species are not widely disseminated, but are also rarely restricted to a single lineage and are instead shared among several, in different combinations (Croucher et al. 2014). It has been suggested that different combinations of accessory loci might be selected in different populations, depending on the overall frequencies of the individual genes, as a result of negative frequency-dependent selection (Corander et al. 2017). At present, this remains a hypothesis without definitive proof.

We must also recognize that a locus might have no wider ecological significance whatsoever. Toxin–antitoxin genes can drive their own acquisition and maintenance, to say nothing of the multitudes of transposable elements, prophage and the like (Wozniak and Waldor 2009). Bacterial genomes are characterized not only by their variable gene content, and the transience of the associations between loci (long for core genes, short for others) but by the divergent selective processes affecting them. In some cases (the core) these are aligned, while in others they are not. Population geneticists who study sexually reproducing eukaryotes are familiar with the notion that the selective interests of different loci in the same genome may differ. The shuffling of genetic information in each generation effectively uncouples the association between all but the closest loci, but even the most frequently recombining bacteria (Arnold et al. 2018) do not approach the state of sexually reproducing eukaryotes. As a result, the overall fitness of a bacterial genome is the product of all the loci making it up. To preserve this overall fitness, it has been proposed that homologous recombination in bacteria is an adaptation to prevent the colonization of the genome with selfish genetic elements, by rapidly replacing them with the homologous region in the ancestral strain, which lacks the additional gene (although this does not explain the notable variation among bacteria in their recombination rates) (Croucher et al. 2016). One of the greatest challenges in providing a satisfying account of bacterial population genetics has been separating the patterns that are the result of selection, from those of linkage.

The question of how the individual loci that make up the accessory or dispensable part of the pangenome, associate themselves with the lineages that are defined by the core component, has come under increasing scrutiny as the numbers of population genome samples have increased. Population genetic models for the core genome specifically developed with bacteria in mind, and capable of handling the various amounts of homologous recombination, are not common. Rarer still are models that explicitly consider the gain and loss of genes from the accessory genome. Although gain and loss of loci is not unknown in eukaryotes, and has been implicated in some major adaptive events (McInerney 2017; Schönknecht et al. 2014) it is nowhere near as extensive and does not have anything like the impact it does in bacteria. Accurate models for such processes are crucial to detect departures from neutrality, and several studies have actually found apparently neutral associations between elements of the pangenome and the core. However there are reasons to think that the sequence variation associated with the accessory genome may produce fundamentally different results from those in population genetics textbooks. For example, if the site frequency spectrum expected under neutral assumptions is extended to allow mutations in loci that can be gained or lost, systematic bias results (Baumdicker 2015; Baumdicker et al. 2012; Collins and Higgs 2012).

Given that the accessory fraction of the pangenome is enriched for loci involved in properties from toxin production, to restriction-modification systems, and surface antigens, to say nothing of drug-resistant genes, it is hard to imagine that it might fit well to a neutral model—in several cases though, it does (Baumdicker et al. 2012; Marttinen and Hanage 2017). This result is hard to accept, and it should be given all we know about the power of selection and the size of bacterial populations. However, it should be appreciated that a multitude of selective scenarios can produce a signal that is hard or impossible to distinguish from neutrality. Study of other metrics may be required to unveil the underlying processes. For instance, the rates with which diverging strains of pneumococcus acquired or lost genes was found to be indistinguishable from neutrality and even to yield good estimates of the population mutation and recombination rates (Marttinen and Hanage 2017). Yet later analysis of the same population, alongside others, was interpreted as strong evidence for negative frequency-dependent selection on the accessory fraction of the genome (see above). What is going on?

A possible explanation lies in the central limit theorem. If an outcome is determined by many independent random variables, each with finite variance, then we expect the result of adding them all together to be a normal distribution. In other words, if the fitness of a strain is the consequence of many independent factors, we might find it appears neutral—the chances of any individual getting into the next generation could be normally distributed around 50:50. This result has been the source of substantial interest in ecology, given that it can be used to show that species abundance distributions (SADs—a common metric for summarizing ecological diversity (McGill et al. 2007) can appear neutral while actually being the result of many non-neutral processes. In the case of bacteria, the fitness effects of genes on the same mobile element may not be independent, however, the effects of multiple mobile elements may similarly approximate to an overall strain fitness not distinguishable from neutrality. Other models from community ecology may be useful in determining the contents of genomes, as well as ecosystems.

Nevertheless, the current consensus in the field is that gene variation directly reflects the ecological niche occupied by the bacteria (Sheppard et al. 2018) and the response to local selective pressures (Cordero and Polz 2014). This may involve the acquisition of antibiotic-resistant genes, as described above, metabolic genes needed to exploit a novel energy source, bacteriocins for microbial warfare, or phage and phage-defense genes involved in predation–prey "paper-rock-scissor" dynamics, as so eloquently described by Corander and colleagues (Arnold et al. 2018). Further, it is suspected that rapid acquisition and dissemination of genes most often occurs as bacterial clones adapt to a novel niche previously occupied by another species (Polz et al. 2013; Popa et al. 2011; Smillie et al. 2011; Vos et al. 2015). An example of this would be the acquisition of IncA/C plasmid by *Vibrio cholerae* introduced to Haiti, a country previously devoid of epidemic cholera for at least 100 years (Carraro et al. 2016) as well as the post-vaccine population of *S. pneumoniae* in the USA, which experienced a significant population shift after the 7-valent pneumococcal conjugate vaccine removed approximately 30% of the pre-vaccine population. Niches themselves are not explicitly segregated, and therefore one does not have to be vacated to then be exploited by a newcomer. Gene flow may occur between sympatric lineages; i.e., habitat borders are not defined by walls or other barriers, and recombination can occur among lineages of a species where habitat space is not clearly demarcated (Marttinen and Hanage 2017). This model explains lineage divergence and population structure among several species, and is important because it highlights that a species requires not only the ability to acquire genes but also the opportunity to do so. Interestingly, it has also been suggested that once a competent species encounters a new niche, it can give rise to noncompetent lineages, providing an advantage when adaptation through gene acquisition is not required and may, in fact, be deleterious (Jorth and Whiteley 2012).

The acquisition of genes is not always beneficial and may, in fact, be deleterious (Vos et al. 2015). Indeed, for every successful lineage that is observed, there are likely several "failed" ecological experiments. Since there is not a clear delineation between the fitness gain and costs of gene acquisition, it may be an oversimplification of the dynamics to ascribe a net-positive or net-negative effect of gene gain and loss. The truth, of course, is somewhere in between and likely varies to different degrees between species. To offset fitness costs and compensate for the acquisition of mobile elements, mutations may arise in core loci and form epistatic relationships with the acquired gene. This has been suggested, for example, in *E. coli*, where nucleotide substitutions in regulatory genes were found to be associated with the acquisition and maintenance of accessory genes (McNally et al. 2016). This dynamic is further supported, in part, by recent findings of epistatic interactions across genome-wide loci among multiple bacterial species (Arnold et al. 2018; Skwark et al. 2017).

There are examples where ecological niches are clearly defined among species and others where the relationships between habitat and organism are obscured. In the *E. coli* study (McNally et al. 2016), ecological adaptation and niche segregation were not observed among isolates collected from humans and animals, while in other species such as *Campylobacter*, this is commonly observed (Sheppard et al. 2011). Methods to investigate gene flow and selection in the context of adaptation and ecology are continually being refined. In some instances, identifying the appropriate system to test ecological hypotheses is the limiting step. An intriguing approach to understanding these associations is not to identify niches, the organisms that inhabit them, and then attempt to resolve the genes associated with adaptation, but instead first assess gene flow and then make predictions about ecology. So-called "reverse-ecology" proposed by Shapiro and Polz seeks to investigate habitat specificity by assessing gene flow and gene-specific sweeps, and has been used to predict ecological differentiation of *Vibrio* spp. in aquatic environments (Hunt et al. 2008; Shapiro et al. 2012; Shapiro and Polz 2014). They demonstrate an example of applying a fresh perspective to an appropriate model system to understanding bacterial ecology.

Taken together, the accumulation of population samples that have been analyzed with modern genomic methods has greatly improved our understanding of the pangenome, and its ecological significance. The totality of loci in a sample includes the essential core, together with a set of accessory loci that have a range of ecological and evolutionary significance: from functional genes with direct relevance to niche such as those described in the reverse ecology approach of Shapiro and Polz, to more selfish elements such as toxin–antitoxin systems. One feature of the current landscape of bacterial genomics that is not often noticed, is that for all the references in the literature to "Whole Genome Sequencing," few studies actually determine the whole, i.e., finished genome including all plasmids. Our current understanding is overwhelmingly based on high-quality draft, not finished, genomes. The emergence of long-read technologies is changing this, and as they improve and become more economical (together with more methods for making hybrid assemblies from short- and long-read data) we may find that our current understanding underestimates the actual quantities of sequence variation in bacteria and that there are short regions under strong selection that accumulate rapid change and are hence hard to assemble from short-read data. Adding these is just one of the exciting directions for research over the next few years, which is sure to improve our understanding of pangenomes and their significance far beyond our current knowledge.

Acknowledgments WPH would like to acknowledge support from U54GM088558 from NIGMS, and helpful discussions with Franz Baumdicker and Claire-Marie MacLean.

References

Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M et al (2018) Weak epistasis may drive adaptation in recombining bacteria. Genetics. https://doi.org/10.1534/genetics.117.300662

Azarian T, Grant L, Georgieva M, Hammitt L, Reid R, Bentley S et al (2016) Pneumococcal protein antigen serology varies with age and may predict antigenic profile of colonizing isolates. J Infect Dis. https://doi.org/10.1093/infdis/jiw628

Azarian T, Grant LR, Arnold BJ, Hammitt LL, Reid R, Santosham M et al (2018) The impact of serotype-specific vaccination on phylodynamic parameters of *Streptococcus pneumoniae* and the pneumococcal pan-genome. PLoS Pathog 14. https://doi.org/10.1371/journal.ppat.1006966

Baumdicker F (2015) The site frequency spectrum of dispensable genes. Theor Popul Biol 100:13–25. https://doi.org/10.1016/j.tpb.2014.12.001

Baumdicker F, Hess WR, Pfaffelhuber P (2012) The infinitely many genes model for the distributed genome of bacteria. In: Genome biology and evolution. Oxford University Press, Oxford, pp 443–456. https://doi.org/10.1093/gbe/evs016.

Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M et al (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet 2:e31. https://doi.org/10.1371/journal.pgen.0020031

Bilek N, Ison CA, Spratt BG (2009) Relative contributions of recombination and mutation to the diversification of the opa gene repertoire of *Neisseria gonorrhoeae*. J Bacteriol 191:1878–1890. https://doi.org/10.1128/JB.01518-08

Blanquart F, Lehtinen S, Lipsitch M, Fraser C (2018) The evolution of antibiotic resistance in a structured host population. J R Soc Interface 15:20180040. https://doi.org/10.1098/rsif.2018.0040

Carraro N, Rivard N, Ceccarelli D, Colwell RR, Burrus V (2016) IncA/C conjugative plasmids mobilize a new family of multidrug resistance islands in clinical *Vibrio cholerae* Non-O1/Non-O139 isolates from Haiti. MBio 7. https://doi.org/10.1128/mBio.00509-16

Chen J, Quiles-Puchalt N, Chiang YN, Bacigalupe R, Fillol-Salom A, Chee MSJ et al (2018) Genome hypermobility by lateral transduction. Science 362:207–212. https://doi.org/10.1126/science.aat5867

Cobey S, Baskerville EB, Colijn C, Hanage W, Fraser C, Lipsitch M (2017) Host population structure and treatment frequency maintain balancing selection on drug resistance. J R Soc Interface 14. https://doi.org/10.1098/rsif.2017.0295

Collins RE, Higgs PG (2012) Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol Biol Evol 29:3413–3425. https://doi.org/10.1093/molbev/mss163

Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD et al (2017) Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nat Ecol Evol 1:1950–1960. https://doi.org/10.1038/s41559-017-0337-x

Cordero OX, Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol 12:263–273. https://doi.org/10.1038/nrmicro3218

Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD (2012) A high-resolution view of genome-wide pneumococcal transformation. PLoS Pathog 8:e1002745. https://doi.org/10.1371/journal.ppat.1002745

Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J et al (2013) Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet 45:656–663. https://doi.org/10.1038/ng.2625

Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP (2014) Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Commun 5:1–12. https://doi.org/10.1038/ncomms6471.

Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD, Fraser C (2016) Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. PLoS Biol 14:e1002394. https://doi.org/10.1371/journal.pbio.1002394

Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV et al (2010) Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol 11:R107. https://doi.org/10.1186/gb-2010-11-10-r107

Feng Y, Chen C-J, Su L-H, Hu S, Yu J, Chiu C-H (2008) Evolution and pathogenesis of *Staphylococcus aureus*: lessons learned from genotyping and comparative genomics. FEMS Microbiol Rev 32:23–37. https://doi.org/10.1111/j.1574-6976.2007.00086.x

Fortier L-C, Sekulovic O (2013) Importance of prophages to evolution and virulence of bacterial pathogens. Virulence 4:354–365. https://doi.org/10.4161/viru.24498

Georgieva M, Kagedan L, Lu Y-J, Thompson CM, Lipsitch M (2018) Antigenic variation in *Streptococcus pneumoniae* PspC promotes immune escape in the presence of variant-specific immunity. MBio 9:e00264–e00218

Griffith F (1928) The significance of pneumococcal types. J Hyg (Lond) 27:113–159

Hollingshead SK, Becker R, Briles DE (2000) Diversity of PspA: mosaic genes and evidence for past recombination in *Streptococcus pneumoniae*. Infect Immun 68:5889–5900. https://doi.org/10.1128/IAI.68.10.5889-5900.2000

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320:1081–1085. https://doi.org/10.1126/science.1157890

Iannelli F, Oggioni MR, Pozzi G (2002) Allelic variation in the highly polymorphic locus pspC of *Streptococcus pneumoniae*. Gene 284:63–71. https://doi.org/10.1016/S0378-1119(01)00896-4

International Working Group on the Classification of Staphylococcal Cassette Chromosome Elements (IWG-SCC) (2009) Classification of staphylococcal cassette chromosome mec (SCCmec): guidelines for reporting novel SCCmec elements. Antimicrob Agents Chemother 53:4961–4967. https://doi.org/10.1128/AAC.00579-09

Jorth P, Whiteley M (2012) An evolutionary link between natural transformation and CRISPR adaptive immunity. MBio 3. https://doi.org/10.1128/mBio.00309-12

Kislyuk AO, Haegeman B, Bergman NH, Weitz JS (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics 12:32. https://doi.org/10.1186/1471-2164-12-32

Lehtinen S, Blanquart F, Croucher NJ, Turner P, Lipsitch M, Fraser C (2017) Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. Proc Natl Acad Sci USA 114:1075–1080. https://doi.org/10.1073/pnas.1617849114

Marttinen P, Hanage WP (2017) Speciation trajectories in recombining bacterial species. PLoS Comput Biol 13:e1005640. https://doi.org/10.1371/journal.pcbi.1005640

Marttinen P, Croucher NJ, Gutmann MU, Corander J, Hanage WP (2015) Recombination produces coherent bacterial species clusters in both core and accessory genomes. Microb Genomics 1. https://doi.org/10.1099/mgen.0.000038

McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK et al (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol Lett 10:995–1015. https://doi.org/10.1111/j.1461-0248.2007.01094.x

McInerney JO (2017) Horizontal gene transfer is less frequent in eukaryotes than prokaryotes but can be important (retrospective on DOI 10.1002/bies.201300095). Bioessays 39:1700002. https://doi.org/10.1002/bies.201700002

McInerney JO, McNally A, O'Connell MJ (2017) Why prokaryotes have pangenomes. Nat Microbiol 2:17040. https://doi.org/10.1038/nmicrobiol.2017.40

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. https://doi.org/10.1371/journal.pgen.1006280

Ochman H, Davalos LM (2006) The nature and dynamics of bacterial genomes. Science 311:1730–1733. https://doi.org/10.1126/science.1119966

Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet 29:170–175. https://doi.org/10.1016/j.tig.2012.12.006

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res 21:599–609. https://doi.org/10.1101/gr.115592.110

Reyes-Robles T, Dillard RS, Cairns LS, Silva-Valenzuela CA, Housman M, Ali A et al (2018) Vibrio cholerae outer membrane vesicles inhibit bacteriophage infection. J Bacteriol 200. https://doi.org/10.1128/JB.00792-17

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH et al (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239:226–235. https://doi.org/10.1016/j.jtbi.2005.08.037

Romero P, Croucher NJ, Hiller NL, Hu FZ, Ehrlich GD, Bentley SD et al (2009) Comparative genomic analysis of ten *Streptococcus pneumoniae* temperate bacteriophages. J Bacteriol 191:4854–4862. https://doi.org/10.1128/JB.01272-08

Schönknecht G, Weber APM, Lercher MJ (2014) Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. Bioessays 36:9–20. https://doi.org/10.1002/bies.201300095

Seed KD, Bodi KL, Kropinski AM, Ackermann H-W, Calderwood SB, Qadri F et al (2011) Evidence of a dominant lineage of *Vibrio cholerae*-specific lytic bacteriophages shed by cholera patients over a 10-year period in Dhaka, Bangladesh. MBio 2:e00334–e00310. https://doi.org/10.1128/mBio.00334-10.

Shapiro BJ, Polz MF (2014) Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol 22:235–247. https://doi.org/10.1016/j.tim.2014.02.006

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G et al (2012) Population genomics of early events in the ecological differentiation of bacteria. Science 336:48–51. https://doi.org/10.1126/science.1218198

Sheppard SK, Colles FM, McCarthy ND, Strachan NJC, Ogden ID, Forbes KJ et al (2011) Niche segregation and genetic structure of campylobacter jejuni populations from wild and agricultural host species. Mol Ecol 20:3484–3490. https://doi.org/10.1111/j.1365-294X.2011.05179.x

Sheppard SK, Guttman DS, Fitzgerald JR (2018) Population genomics of bacterial host adaptation. Nat Rev Genet 19:1–17. https://doi.org/10.1038/s41576-018-0032-z

Shintani M, Sanchez ZK, Kimbara K (2015) Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. Front Microbiol 6:242. https://doi.org/10.3389/fmicb.2015.00242

Skwark MJ, Croucher NJ, Puranen S, Chewapreecha C, Pesonen M, Xu YY et al (2017) Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. PLoS Genet 13:e1006508. https://doi.org/10.1371/journal.pgen.1006508

Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480:241–244. https://doi.org/10.1038/nature10571

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. Proc Natl Acad Sci 102:13950–13955. https://doi.org/10.1073/pnas.0506758102

Vos M, Hesselman MC, te Beek TA, van Passel MWJ, Eyre-Walker A (2015) Rates of lateral gene transfer in prokaryotes: high but why? Trends Microbiol 23:598–605. https://doi.org/10.1016/j.tim.2015.07.006

Waldor MK, Friedman DI (2005) Phage regulatory circuits and virulence gene expression. Curr Opin Microbiol 8:459–465. https://doi.org/10.1016/J.MIB.2005.06.001

Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D et al (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA 99:17020–17024. https://doi.org/10.1073/pnas.252529799

Wiedenbeck J, Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35:957–976. https://doi.org/10.1111/j.1574-6976.2011.00292.x

Wozniak RAF, Waldor MK (2009) A toxin-antitoxin system promotes the maintenance of an integrative conjugative element. PLoS Genet 5:e1000439. https://doi.org/10.1371/journal.pgen.1000439

Wozniak RAF, Fouts DE, Spagnoletti M, Colombo MM, Ceccarelli D, Garriss G et al (2009) Comparative ICE genomics: insights into the evolution of the SXT/R391 family of ICEs. PLoS Genet 5:e1000786. https://doi.org/10.1371/journal.pgen.1000786

Young JPW (2016) Bacteria are smartphones and mobile genes are apps. Trends Microbiol 24:931–932. https://doi.org/10.1016/j.tim.2016.09.002

Bacterial Microevolution and the Pangenome

Florent Lassalle and Xavier Didelot

Abstract The comparison of multiple genome sequences sampled from a bacterial population reveals considerable diversity in both the core and the accessory parts of the pangenome. This diversity can be analysed in terms of microevolutionary events that took place since the genomes shared a common ancestor, especially deletion, duplication, and recombination. We review the basic modelling ingredients used implicitly or explicitly when performing such a pangenome analysis. In particular, we describe a basic neutral phylogenetic framework of bacterial pangenome microevolution, which is not incompatible with evaluating the role of natural selection. We survey the different ways in which pangenome data is summarised in order to be included in microevolutionary models, as well as the main methodological approaches that have been proposed to reconstruct pangenome microevolutionary history.

Keywords Pangenome · Bacterial microevolution · Evolutionary model · Recombination · Duplication · Deletion · Gene content · Ancestral state reconstruction · Reconciliation

1 Atomic Events in Bacterial Microevolution

Bacterial microevolution is the study of the evolutionary forces that shape the genetic diversity of a natural population of bacteria. This evolutionary process takes place as a result of the genetic changes happening within each of the genomes of the bacterial cells constituting the population. Over time, these changes are amplified or weakened by the effects of both genetic drift and natural selection.

F. Lassalle
Department of Infectious Disease Epidemiology, Imperial College London, London, UK

X. Didelot (✉)
School of Life Sciences and Department of Statistics, University of Warwick, Coventry, UK
e-mail: xavier.didelot@warwick.ac.uk

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_6

Genetic drift represents the evolution caused by the death and birth of cells in the bacterial population, and it acts at random on all genetic variants (Charlesworth 2009). The effects of genetic drift are higher when the population size is small, and so it could be thought given the large number of cells in bacterial populations that genetic drift would be weak. However, bacterial populations sometimes go through punctual bottlenecks during which genetic drift has a large effect, for example during transmission of pathogens from one host to another (Didelot et al. 2016). It is also believed that the strong structure of bacterial habitat, sometimes at the microscale can lead to much smaller effective population sizes than intuition suggests (Vos et al. 2013). Natural selection on the other hand acts in a nonrandom fashion, amplifying some variations and suppressing others, and is a very potent evolutionary force in shaping the diversity of bacterial species (Petersen et al. 2007; Buckee et al. 2008; Pepperell et al. 2013).

The genetic changes occurring on a single bacterial cell can be classified into mutation and recombination events, and the events of interest differ whether the focus is on the core genome (the regions shared by all genomes in the population) or the accessory genome (the regions that are found in some but not all of the genomes). As far as the core genome is concerned, the main type of mutation is the point mutation, whereby a single nucleotide is replaced, and the main type of recombination is called homologous recombination, in which a relatively short fragment of the genome is replaced with a homologous fragment coming from another bacterial cell (Didelot and Maiden 2010). There are three biological mechanisms that can lead to homologous recombination, namely conjugation (where two bacterial cells come in contact so that DNA can be transmitted from donor to recipient), transduction (where a phage acts as vector from donor to recipient) and transformation (where naked DNA is picked up by the recipient from the environment, possibly following the death of the donor cell) (Thomas and Nielsen 2005). But since their outcomes are hard to distinguish this diversity of mechanisms is usually ignored in evolutionary models of homologous recombination.

Point mutation and homologous recombination events clearly act on the evolution of the accessory genome in the same way as they do for the core genome. However, they do not change the genetic content of core genomes. There are two types of endogenous mutations that can alter the genetic content of a genome, duplication and deletion, and they can be thought of as opposite forces, with the former increasing the number of copies of a gene by one and the latter decreasing it by one. Finally, the accessory genome is also subject to non-homologous recombination, where a bacterial cell imports a DNA fragment from another cell and inserts it in its genome, without overwriting a previously existing homologous fragment (Ochman et al. 2000). Non-homologous recombination is often called lateral gene transfer or horizontal gene transfer, and in this chapter we will be using these three terms interchangeably. It should be noted, however, that this terminology is not always consistently used in the literature, with some authors using the term horizontal gene transfer to refer to both homologous and non-homologous recombination.

The three biological mechanisms mentioned above for homologous recombination (conjugation, transduction, and transformation) can lead to non-homologous recombination and once again, it is helpful when studying the bacterial microevolution of the accessory genome to set aside the mechanism at play. Likewise, genetic duplication and deletion can have multiple causes that we will not explore. It should in fact be noted that even though it is useful to present and study them as separate, the atomic evolutionary events briefly described above are not biologically independent (Lawrence 1999; Everitt et al. 2014; Oliveira et al. 2017). For example, a single event of recombination could involve the replacement of some genes (homologous recombination), the insertion of new genes (non-homologous recombination) and the loss of some other genes (genetic deletion).

Furthermore, non-homologous recombination can sometimes be duplicative if the newly imported material is homologous to a sequence found somewhere else in the genome. In this case, the number of copies of the genes concerned is increased by one, as in a duplication event. The evolutionary distance between donor and recipient of such a non-homologous recombination event is then a crucial factor: if this distance is small the effect is similar to a duplication event, which can be seen as a transfer event where recipient and donor are the same organism. If distance between organisms is high, then the difference between the newly imported copy and the copy already present will likely be high too, providing a clear sign that duplication was not involved. This situation is analogous to the detection of homologous recombination in the core genome, where events from a closely related source do not leave a trace, or perhaps involve just a single substitution in which case they are undistinguishable from point mutation (Didelot et al. 2010).

2 Neutral Phylogenetic Framework of Bacterial Pangenome Microevolution

2.1 Challenges with a Comprehensive Model

The microevolutionary events that act on the bacterial pangenome, as briefly described above, can be combined into an evolutionary model of how the pangenome evolves over time. Let us consider a comprehensive model, which would account for the whole population of bacterial cells, including the fact that cells die and reproduce over time (so that genetic drift is included) and that various selective pressures are exerted. In this model, the genome of each cell is affected by various mutation and recombination events, all of which happens at a certain rates over time for each cell. All the rates involved in this model (birth and death of individuals, selection for specific variants and various evolutionary events) would not be assumed to be constant, but would be allowed to vary over time. This model falls in the class of forward-in-time models, due to the fact that it considers evolution as it unfolds over time, and famous examples of such models in the general population genetics literature are the Wright–Fisher model (Fisher 1931; Wright 1931) and the Moran model (Moran 1958). Figure 1 illustrates such a forward-in-time model of pangenome evolution.

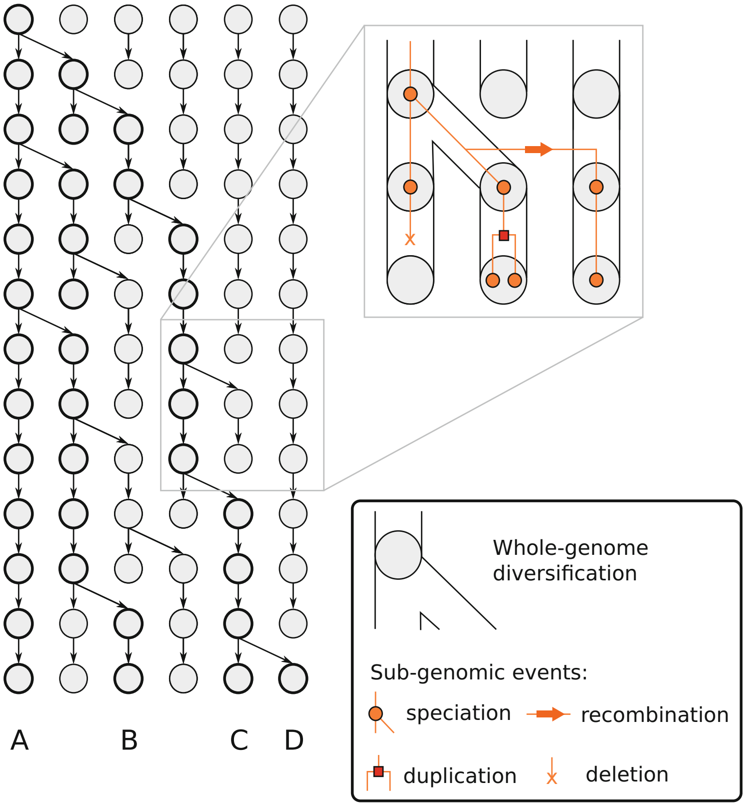


Fig. 1 Illustration of the forward-in-time evolution of a bacterial population and its pangenome. At each time step, an individual is removed and another gives birth as in a standard Moran model. Furthermore, at each time step the accessory genome of each individual may evolve via deletion (orange cross), duplication (red square) and recombination (orange arrow)

The idea of this comprehensive model is to replicate exactly the processes that we know occur in nature, so that it is of the highest possible realism. However, a comprehensive model would also integrate the diversification process of the whole community of microorganism found at a given spot, with the impact of their biotic interactions and genetic exchange, but most importantly, of the competitive process leading to natural selection of the fittest. Such level of description of natural processes would render the model impossible to use, and that is why it is not found in the literature. It is educative though to ask ourselves why this model is unusable, as this will guide us towards more practical models that feature some of these ideal properties.

The first problem with this comprehensive model is computational: it would require very large amounts of computer memory to store the state of a population at a single time point, even much more so to track its evolution over time, and an equally impossible amount of computer power to consider the evolutionary events happening to all members of the population. But even more importantly, there is statistical problem with the comprehensive model, in the sense that there are too many unknown quantities involved, for which we would not be able to take even a very rough guess at what their value might be. Therefore, even if the computational problems could be overcome, and analysis conducted under the comprehensive model, the results would be worthless since the quantities to be estimated would be unidentifiable. Simplifications will therefore have to be made to reach a model that has practical use, with the best model being not the most comprehensive one, but the one that achieves the best trade-off between biological realism on one hand, and computational and statistical considerations on the other hand.

Beyond the degree of complexity of a model and the search for a trade-off between computational efficiency and model realism, models may rely on different conceptual formalisation of bacterial genomes and their evolutionary process. These different concepts will generate different approaches and methods that are in general complementary. We will thus present different elements of phylogenetic models of pangenome evolution, which flavours may be combined to provide a variety of practical models.

2.2 Analysing Selection Based on Neutral Models

Perhaps the greatest challenge posed by the comprehensive model above would be its attempt at encompassing the role of natural selection. As previously mentioned, it is clear that natural selection plays a crucial role in shaping the microevolution of the pangenomes of bacterial populations, but the effect of this force is different for all genes or nucleotidic site and their allelic variants, may vary significantly over time, and be different for different segments of the populations, for example if some lineages are adapted to a certain environment. Such adaptation of a lineage will involve many traits distributed in the pangenome of that population, and new mutation arising in this background might interact with it; this leads to complex epistatic (i.e. non-additive fitness) interactions between genomic traits, affecting the probability of selecting new genetic variants in one or another genomic background—a process that could add infinite degrees of complexity to the exhaustive model. Model design can, therefore, be greatly simplified by considering no effect of selection, or in other words neutrality of evolution.

Even if the role of selection is not explicitly included in a model, it does not mean that analyses based on this model are completely uninformative about selection. Neutral evolutionary models provide a framework to search for evidence of natural selection. This can be achieved formally by contrasting observed patterns in compared genomic data to expectations under neutral models. Another approach to detect selection is to fit a neutral model to genomic data, having heterogeneous parameters to describe the evolution process of each species lineage and/or gene family; outlier species or genes with ‘abnormally’ high or low parameters can provide a clue to non-neutral processes taking place. Similarly, the identification of historical changes of processes (e.g. acceleration or slowing down of diversification rates) in the scenario of pangenome evolution can provide strong clues of selection affecting the species lineage or gene under focus (Boussau et al. 2004; David and Alm 2011; Lassalle et al. 2017).

This approach is similar to the way that the role of selection is being investigated in the core genome. In this more frequently explored setting, a typical pipeline (Hedge and Wilson 2016) involves reconstructing a phylogenetic tree, classifying substitutions in terms of whether they are synonymous or not and estimating evolutionary rates so that selection tests can be applied based on variations in the rates of synonymous and non-synonymous substitutions along the genome (Wilson and McVean 2006; Castillo-Ramírez et al. 2011) or between populations (McDonald and Kreitman 1991; Vos 2011). In this popular approach, the evolutionary models used to build the phylogenetic tree and reconstruct substitution events are purely neutral, but still lead to invaluable insights into the natural selection process.

2.3 Phylogenetic Approach

A neutral version of the comprehensive model described above would still be impossible to use in practice. A major difficulty is that it considers the evolution of the population forward-in-time, so that every single cell in the population has to be included. However, any dataset we may have available for analysis will only include a small fraction of the population, sampled typically at a single time point (or a few recent time points in the best-case scenario). However, considering the evolution of the whole population over time can appear wasteful, since most cells in the past would not have had any descendants surviving in the present-time sample. A much more tractable approach is therefore to only consider the genealogical process of the sampled genomes, which is a backward-in-time process. Under relatively mild assumptions, and without introducing too much approximation, this genealogical process can be described without reference to the whole forward-in-time process. In particular, the coalescent model (Kingman 1982) describes the genealogical process of a population following either the Wright–Fisher or the Moran model of forward-in-time evolution. Extensions of the basic coalescent model have been derived to deal with fluctuating population size (Griffiths and Tavare 1994), homologous recombination (Griffiths and Marjoram 1997), which for bacteria is analogous to gene conversion (Wiuf and Hein 2000), and many other forms of relaxation of the assumptions ruling the evolutionary process (Donnelly and Tavare 1995; Nordborg 2001; Rosenberg and Nordborg 2002).

Considering this genealogical process, and the ability to reconstruct it with relatively high accuracy from genome sequences, is pivotal to lead to a usable model of pangenome evolution. A simple approach is to focus on core genome elements and apply a standard phylogenetic method typically based on maximum likelihood or Bayesian inference under an evolutionary model of neutral point mutations (Yang and Rannala 2012). Bacteria reproduce clonally and most species recombine relatively rarely (Vos and Didelot 2009; Yang et al. 2018), so that this simple approach can often be sufficient for our purposes. Phylogenetic methods have also been developed that can account for the effect of homologous recombination while still reconstructing a single tree (Didelot and Wilson 2015; Croucher et al. 2015). Methods that attempt to reconstruct a graph of ancestry rather than a single tree are superior in principle, but rarely used in practice due to their high computational cost (Didelot et al. 2010; Vaughan et al. 2017).

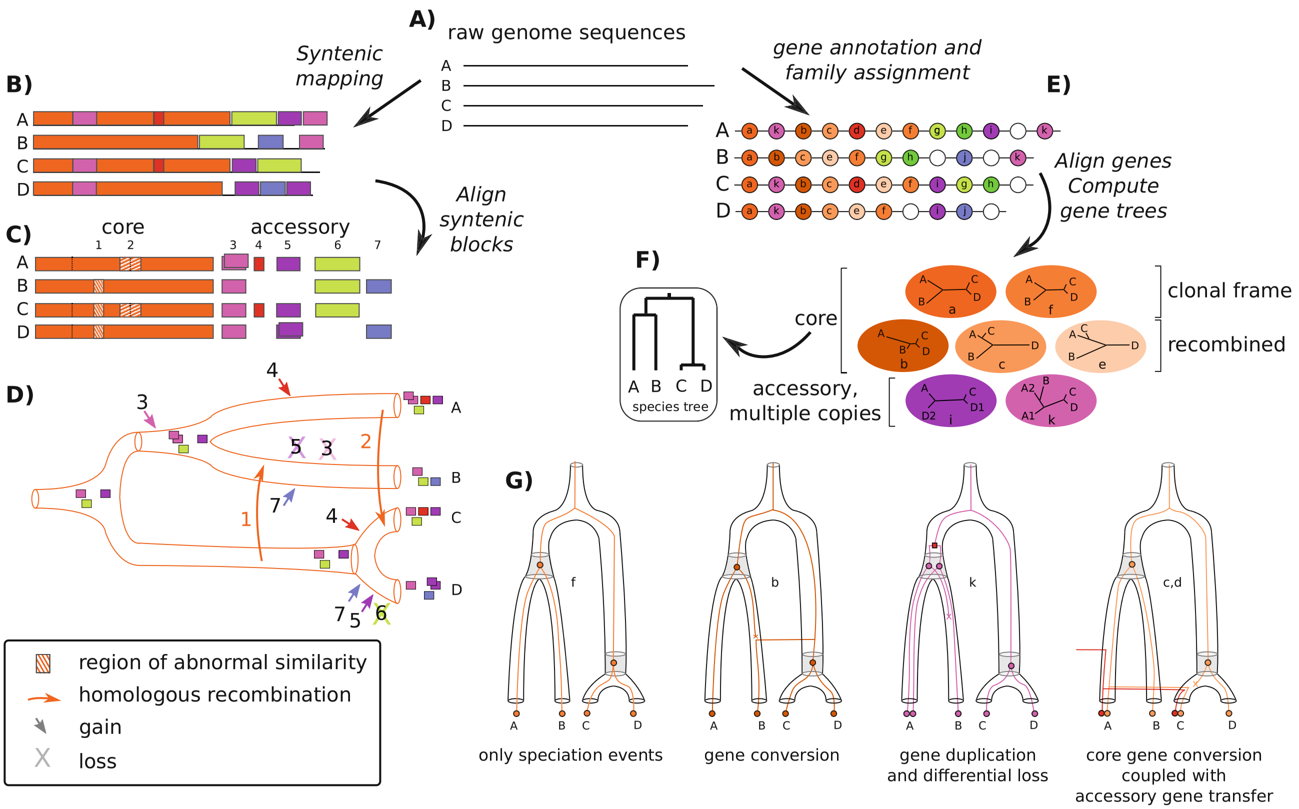
In the context of a phylogenetic tree reconstructed from the core genome, we can consider the events of duplication, deletion and non-homologous recombination that shape the accessory genome. These events happen on the branches of the phylogeny at certain rates that may vary over time and lineages. Notwithstanding such remaining complexities, a phylogenetic model of bacterial pangenome microevolution represents a practical approach relative to the comprehensive forward-in-time model. The events that affect the accessory genome are relatively rare, which results in a strong phylogenetic inertia of genome gene contents, i.e. a large correlation between gene contents and core genome-based diversity (Konstantinidis et al. 2006; Kislyuk et al. 2011). Ignoring this effect would lead to strong misinterpretation of gene distribution patterns, especially in case of a diversity bias in genome sampling, e.g. when surveying a pathogen epidemics where clusters of closely related strains occur. Modelling the pangenome evolution within a phylogenetic framework where evolution of the gene content takes place along the genealogical tree avoids such pitfalls, in the same way as a phylogenetic framework avoids false conclusions to be reached when performing bacterial genome-wide association mapping (Collins and Didelot 2018).

3 Description of Pangenome Data for Inclusion in Microevolutionary Models

3.1 *Units of Pangenome Evolution*

In order to describe further the existing models of pangenome microevolution, it is necessary to consider the unit in which the pangenome is being described. Figure 2 illustrates the different approaches that have been used for that purpose. The ideal starting point would be a complete sequence of each genome of interest, but this is rarely available due to repeat regions in the genomes that obscure the exact ordering of sequences along the genomes, at least based on short read sequencing. For that reason, the most frequently used data is a de novo assembly of each genome, which can be performed, for example using Velvet (Zerbino and Birney 2008) or SPAdes (Bankevich et al. 2012). This results a set of genomic regions called contigs, which occur in an unknown order either on chromosomes or on plasmids.

A first approach considers a genome alignment, where every part of a genome is assigned to a syntenic block—segments of genome sequences that are all



- ▨ region of abnormal similarity
- → homologous recombination
- ↘ gain
- X loss

Fig. 2 (**a**) Datasets of raw genome sequence assemblies can be dealt with different approaches. (**b**) Whole-genome alignment of the compared genomes. Extensive homologous regions (syntenic blocks) are mapped in different colours. Orange indicate the core genome, the other colours indicate accessory syntenic blocks; the absence of a colour block indicate genome-specific (orphan) sequence. (**c**) From there, core and accessory genomes can be treated separately. An ungapped core genome alignment is produced (left hand, orange blocks; dotted vertical lines indicate insertion point of accessory material). Regions of abnormal sequence similarity (with respect to genome-wide divergence) are identified (hatched boxes) as clues of homologous recombination events. Presence of accessory genomic blocks is counted in each genome, an information that can be easily stored in a numeric matrix. (**d**) The clonal genealogy (orange tubes) is computed from the core genome alignment purged of recombined regions. Homologous recombination events and gain/loss of accessory genome blocks can be inferred by confronting this tree to the information gathered in (**c**). (**e**) Another approach consists of annotating the genes present in the genomes (e.g. protein-coding sequences) and to classify them onto homologous gene families based on sequence similarity. (**f**) The occurrence pattern of each family allow to classify them as core (when a single copy of the gene is present in every genome) or accessory. Aligning gene family sequences allow to infer gene-specific phylogenies, or gene trees. Core gene trees can be used to generate a consensus tree representative of the core genome history (species tree). Differences of topologies between the species tree and the individual core gene trees reveal genes subject to homologous recombination. (**g**) Ancestral states and evolution scenarios can be reconstructed for each gene family, through reconciliation of the gene family tree with the species tree under a model with events of gene duplication, transfer and loss. Sample scenarios for families *f*, *b*, *k*, *c* and *d* are presented, with the reconciled gene tree (colour line) embedded in the species tree (tube); shaded areas indicate events of whole-genome diversification, where gene speciation events occur. Note that simpler gene gain/loss scenarios and ancestral gene content reconstruction can be reconstructed based on gene family occurrence data using the method depicted in (**d**) homologous and can be aligned (Fig. 2b, c). These sequence segments can have boundaries falling anywhere in the genome, notably between or within protein-coding sequences, and the often span many genes. While this is a flexible view of pangenome evolution that is probably the most realistic—evolving genomes ignore human annotation of functional elements—it may be cumbersome to implement with a growing number of compared genome. Indeed, every genome added to the dataset may result in the breakage of a syntenic block into several parts due to insertion, deletions, or rearrangements in one of the homologous genome segments. Homologous genome segments can themselves be difficult to align at the nucleotide level when they include fast-evolving genome regions. For these reasons, even the best software for this task such as progressiveMauve (Darling et al. 2010) or MUGSY (Angiuoli and Salzberg 2011) can only deal with between 10 and 100 genomes, depending on how diverse they are. This first alignment approach works best on the well-conserved parts of pangenomes, i.e. the core genomes and possibly large conserved accessory regions of the genomes. This partial sampling of genome sequences is practical because it allows to represent the homology between genomes as a concatenated alignment of all these syntenic blocks, which amounts to a representative map of the genome. Alternatively, a representative whole-genome alignment can be obtained by mapping all homologous sequences in compared genomes to the genome of a reference isolate, using, for example MUMmer (Kurtz et al. 2004). This can result in reference-biased representation, which may be avoided by restricting the alignment to the core genome.

A second approach focuses on genes, or more specifically on families of homologous genes. These are usually defined based on sequence similarity and restricted to protein-coding sequences, even though it can be applied to conserved intergenic sequences as well (Fig. 2e). In this representation, rather than a reference whole-genome map, we consider independent gene families, which members need not be localised in a genome. The diversity of the gene family can conveniently be represented with a phylogenetic tree based on all nucleotide positions in the aligned genes, which allows the estimation of statistical support (Fig. 2f). This information can, in turn, be used to inform the ancestral reconstruction of genome gene content (as discussed below). There are several ways in which this gene family content identification can be performed. If a representative from each family is known in advance, similarity search tools like BLAST (Altschul et al. 1997) can be used to search them in each genome, and, for example BIGSdb automates this process (Jolley and Maiden 2010). Alternatively, each de novo assembled genome can be annotated separately, using, for example Prokka (Seemann 2014) or RAST (Aziz et al. 2008). Homologs can then be identified by using a combination of similarity search between genes from different genomes (e.g. with BLAST) and similarity network analysis, as implemented, for example in the software OrthoMCL (Li et al. 2003), with integrated pipelines implemented in software like Roary (Page et al. 2015) and MMseqs (Steinegger and Söding 2017).

A consequence of this distinction is the way genetic exchange between genomes is considered. In the nucleotide-centred vision, genetic exchange will result either in the replacement of a region (homologous recombination) or the insertion of a sequence at a defined position in the genome map (non-homologous recombination) (Fig. 2c). Similarly, a genetic duplication event will consist in recopying a segment of genome sequence and inserting it next to its template (tandem duplication) or away from it. Homologous recombination events can be evidenced based on a scan of the genome map, looking for increased or decreased sequence similarity (or phylogenetic relatedness) between compared genomes along the genome map (Didelot et al. 2007). Non-homologous recombination and duplication events consist of insertion events and are simply evidenced by some region being only represented in some genomes in the alignment—the others featuring a large 'gap', or long string of missing characters. Distinguishing non-homologous recombination from duplication events can be tricky: even comparing the inserted segment to the rest of the genome and finding a similar region is not conclusive that it would be the duplication template (or copy) of the studied region. Such a pattern could also result from a recombination with a related organism leading to the insertion of genetic material that had homologous counterpart already residing in the recipient genome. Not finding a similar region is not conclusive of the insertion resulting from a non-homologous recombination event either, as an ancient duplication followed by a loss (or deletion) in the compared lineage may result in the same pattern. The answer to this conundrum is modelling of the possible sequences of events, or scenarios, and determine the most likely based on patterns of sequence divergence.

In the second approach centred on genes, the exchange of genetic material is made most evident in the phylogeny of genes, or gene tree, because the gene from the recipient will be more closely related to genes from the donor than to genes from closely related species. In this context, the event is rather called horizontal gene transfer, in opposition to vertical evolution, which would have resulted in the 'normal' clustering of genes from closely related species. Again, this representation ignores the locus where genes sit, and it is therefore not straightforward to know from the gene tree whether the horizontal gene transfer event resulted in the replacement of a resident sequence or in the addition of a new one.

There are also other evolving units that can be considered as the basis of pangenome microevolution modelling, including conserved protein domains or short sequences of a constant length, which are also known as words, features, or k-mers (Sims and Kim 2011; Sheppard et al. 2013b). Some units may seem more natural than others from a theoretical point of view, but in practice all units have pros and cons, and the choice of unit is guided by the evolutionary resolution required by each pangenome investigation.

3.2 Granularity of Homologous Groups

When modelling the pangenome diversity with homologous gene families, a further distinction can be done on which homologous link to consider clustering genes into families. A popular approach is to consider orthology relationships. In theory, genes are orthologous when they are related only by events of speciation (i.e. diversification of the whole genome), not by duplication of horizontal gene transfer. Because the true course of gene diversification events is unknown, we must rely on practical definitions of orthology. This theoretical definition implies that two orthologues cannot occur in the same genome. A usual criteria is thus to look for the bidirectional best hit (BBH) in a similarity search of all the genes in a genome against all of the genes in another genome pairwise genome (Altenhoff and Dessimoz 2009).

This pairwise relationship can be used to build a network of genes covering the whole pangenome dataset, where cliques (groups where the found relationship is transitive among members) are recognised as clusters of orthologous genes or COGs (Tatusov et al. 1997). Using this practical definition, it is straightforward to classify any gene into a cluster, many of which will however be clusters of genes on their own: orphan genes with no homologues, but also those resulting from a recent transfer or duplication. By construction, these COGs can only be absent or present in a single copy in a genome, which proves very convenient for representing the distribution of genes in the pangenome by a genome-to-COG binary matrix filled with zeros and ones. This representation can be handled by many simple methods that model the transition between these binary states over the tree of the genomes, i.e. events of gene gain and loss (Fig. 3a) (Mirkin et al. 2003). This approach has been widely used, but suffers from its stringent definition that leaves many homologous genes out of COGs under scrutiny, which may strongly flaw the inferred ancestral genome gene contents and the derived conclusions on ancestral functional repertoires.

Instead, it is possible to consider a whole family of homologues, which distribution of the family in genomes can again be represented in a matrix of counts, where this time values range from zero to any integer number. Models of pangenome evolution can account for this multiplicity of gene copy number by invoking extra gene gain events (Fig. 3b) (Csurös 2008; Csurös and Miklós 2009). The nature of these gain events—duplication or horizontal gene transfer—is, however, not inferred as it fundamentally requires to know the phylogenetic relationships between genes within a homologous family.

3.3 Linkage of Genes and Syntenic Blocks

Notwithstanding the type of evolving unit considered (aligned genome segment or gene family), all units are usually considered to evolve independently on the phylogeny. This is, however, not always realistic given the high linkage disequilibrium found in bacterial genome—the non-independence of physically linked characters in evolution, a consequence of their clonal mode of reproduction.

Linkage can be introduced in a pangenome evolution model by specifying the location of genes on a genomic map. The evolution of genes on the map is then modelled through events of insertion, deletion and rearrangement. This map can relate the absolute position of genes in contemporary genomes (i.e. with nucleotide site coordinates) by chopping all genomes in the dataset into syntenic blocks, where

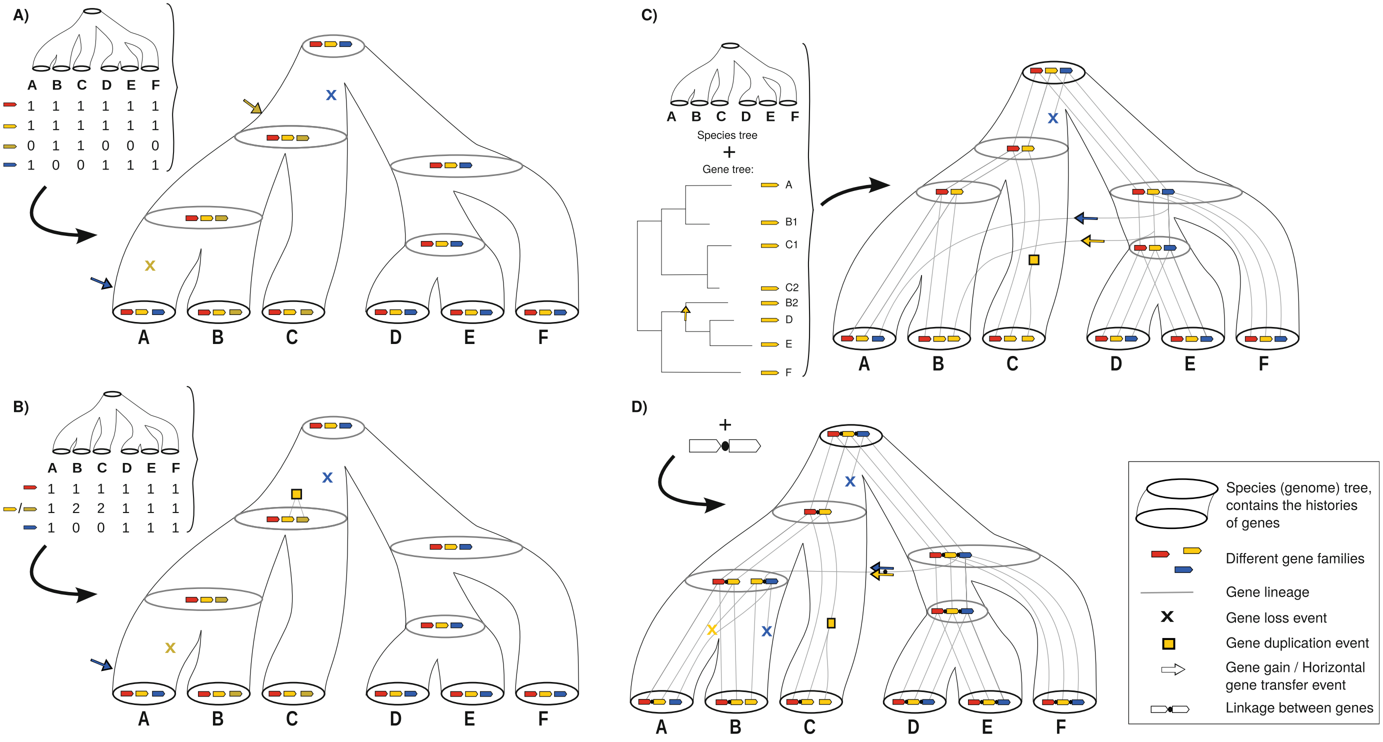


Fig. 3 Schematic representation of scenarios of gene family evolution, inferred with varied levels of information input. (**a**, **b**) Phylogenetic profile: only the information of presence/absence of gene families is used to infer the scenarios of evolution. (**a**) Genes are classified into orthologous groups; (**b**) genes are classified into homologous groups, retaining the information that yellow and brown gene families are related, the emergence of the latter being explained by a duplication of the former. (**c**) Gene tree/species tree reconciliation: the information of topology of the gene tree is used, revealing a more complex scenario, with an additional transfer. (**d**) Synteny-aware reconciliation: the information of linkage of genes (synteny) is used, suggesting that apparently independent transfer events may have happen as one joint event

homology and overall gene order between genomes is conserved (Vallenet et al. 2006; Darling et al. 2010). However, as mentioned above, the larger the genome sample, the more syntenic blocks will split and shrink. Based on such genome maps, the history of each syntenic blocks can be estimated, describing the ancestral events of pangenome evolution. Even though in theory the map evolves over time due to genome rearrangements (Darling et al. 2008), in practice the maps are assumed to be constant in order to allow to focus on fine-grained changes within the syntenic blocks. This assumption is commonly made, for example when investigating homologous recombination in the core genome (Didelot et al. 2010).

Another option is to map the relative position of smaller evolving units (usually gene families) in each genome of the dataset. Such a relative map can be represented by a matrix of presence or absence of a direct adjacency between genes in a given genome, contemporary or ancestral. This more abstract representation allows the use of incomplete data, such as draft genome assemblies, where the physical linkage of sequences is not fully or not unambiguously documented. The evolution of gene neighbourhood is modelled by invoking events of creation and breakage of adjacencies between neighbour genes, thereby modelling any insertion, deletion and rearrangement. Ancestral state reconstruction (see below) is then undertaken, by estimating a genome map at each ancestral node of a species phylogeny (Fig. 3d) (Bérard et al. 2012; Patterson et al. 2013; Duchemin et al. 2017). These models are, however, quite heavy computationally and may become overwhelmed by large structural diversity in the dataset.

4 Methodological Approaches to the Reconstructing Pangenome Microevolution

4.1 Ancestral State Reconstruction

The inference of ancestral genomes and corresponding gene gain and loss scenarios can be a complex and computationally intensive task, but it can also be simplified to the point that it becomes almost straightforward if the research questions are relatively simple. For example, using profiles of gene presence/absence in genomes and a phylogenetic tree as only input, ancestral state reconstruction can be applied to infer in which internal nodes of the tree the genes were present, and therefore on which branches the genes were gained and lost. For a general review on ancestral state reconstruction, see Joy et al. (2016). One of the simplest and most popular approach is to perform a parsimonious reconstruction, where the number of gain and loss events is minimised without the need to estimate any parameter (Mirkin et al. 2003). In practice, this is more or less equivalent to performing maximum likelihood inference under a model in which gain and loss happen at the same small rates. However, probabilistic modelling of state evolution has the interesting property to integrate over several possible scenarios. Even a maximum likelihood point estimate of the presence of a gene at a given ancestral node will therefore consist of a non-binary probability, a nuanced result allowing to consider the uncertainty in the ancestral reconstruction (Pagel 1999). A similar Bayesian approach is stochastic character mapping (Huelsenbeck et al. 2003), which consists in sampling gain and loss histories from their posterior probability distribution via a Monte Carlo method.

Ancestral state reconstruction is particularly well suited to analyses focused on specific genes rather than the whole pangenome, for example analysing the gain and loss of pathogenicity genes (Dingle et al. 2014) or resistance genes (Ward et al. 2014). It can also be applied more generally to all genes in a pangenome, and the rates of gain and loss would typically be assumed to be equal meaning that the genome size is at equilibrium (Touchon et al. 2009). Alternatively, the reconstruction can be based on genomic elements known to be gained and lost in one block, such as bacteriophages, plasmids, and integrative conjugational elements (Zhou et al. 2013). This represents one simple way of dealing with the linkage of genes mentioned previously, although at the cost of potentially losing information about the gene content evolution of the genomic elements assumed to be perfectly linked. At the other end of the spectrum, the reconstruction can be based on smaller elements than genes, for example k-mers, but in this case it becomes vital to relax the assumption of a fixed clock on gain and loss, for example using a local clock model (Didelot et al. 2009) as illustrated in Fig. 4. This technique has been applied to the pangenomes of *Escherichia coli* (Didelot et al. 2012) and *Campylobacter jejuni* (Sheppard et al. 2013a), showing in both cases a strong relationship between evolution of the accessory genome via gain and loss events and evolution of the core genome via homologous recombination.

An important drawback of ancestral state reconstruction methods is that they ignore the nature (recombination or duplication) and origins (recombination donor) of gene gain events, which can yield partial and inaccurate scenarios when the true history is complex, especially with many recombination events. In particular, the exploitable signal from a profile of gene presence/absence in extant genomes are quickly saturated when several gene copies coexist in a genome, and likely descend from separate past events. This issue can sometimes be tackled by defining strict families of orthologs, where every gene is present in one copy or none, but at the cost of losing the information on evolution of homologues. Ancestral state reconstruction could also in principle be applied to data on family of homologues, where each genome can contain zero, one or more copies of a gene. This would require to fit a ladder model similar to the ones used when analysing microsatellite data (Ohta and Kimura 1973; Wilson and Balding 1998). This approach is difficult in practice because bacterial accessory gene families of interest have often too complex histories to reliably infer orthologous groups and have high gain and/or loss rates that quickly saturate signals. It has, however, been applied successfully in studies where genomes of single representatives from fairly distant species were compared, thus ignoring the ‘messy’ variation introduced by within-population evolution (Csurös and Miklós 2009).

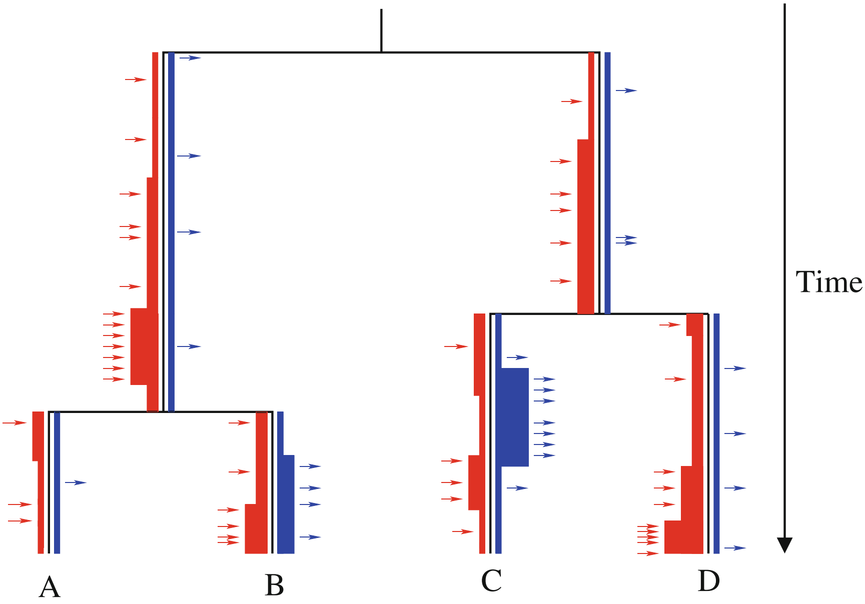


Fig. 4 Illustration of a pangenome gain and loss model with local clock. The clonal genealogy is shown in black. The width of the red block on the left of the branches is proportional to the rate of gain. Similarly, the blue block on the right of each branch represents the rate of loss. Both the gain and loss rates occasionally change over time. Individual gain events are represented by red arrows, and individual loss events are represented by blue arrows

4.2 Phylogenetic Reconciliation

To deliver more informative scenarios of evolution, it is necessary to know the origin of gene gains, which effectively means to know the relationship between observed genes. Gene tree versus species tree reconciliation methods compare the topologies of phylogenetic trees built from individual gene sequences against a reference species tree (Maddison 1997). In the context of pangenome analysis, the species tree is a phylogenetic tree reconstructed from the whole of the core genome. Species and gene trees often have inconsistent topologies, which could happen by chance, especially since the gene tree typically has limited statistical support, or may be the result of evolutionary events affecting the history of the gene relative to the clonal history. Reconciliation methods intend to explain the significant topological discords by events of gene duplication, transfer, or loss (Szollosi et al. 2015). Figure 3c illustrates the principles behind reconciliation methods. Practically, both trees are annotated with the inferred events, such that there is a full agreement on the course of events, from the root of the gene lineage to the contemporary distribution of genes in genomes—thus reconstructing the path of evolution and diversification of genes in the clonal frame of genome evolution. As a result, this approach allows to explicitly determine the donors and recipients of transferred genes, or the ancestor in which a gene was duplicated. Methods for pangenome reconciliation analysis have been proposed based on parsimonious reconstruction (Abby et al. 2010; Jacox et al. 2016) and probabilistic models (Szollosi et al. 2012, 2013).

The ancestral state reconstruction approach and the reconciliation approach have a lot in common, and the latter can be thought of as a natural extension of the former when observation is not limited to presence or absence or number of copies of a gene, but also includes the phylogenetic relationships between genes from separate genomes. Reconciliation methods are therefore superior in the sense that they exploit more of the available signal, but they are also much more challenging to implement computationally and have so far been limited to analysis of a handful of genomes. Ancestral state reconstruction methods are currently more popular but we predict that reconciliation methods will become increasingly widespread in the near future with the development of more effective statistical methods. Beyond the study of the atomic events whereby the pangenome evolves, both methods allow to infer ancestral states in hypothetical ancestors, or in other words to reconstruct ancestral genomes. Doing so, one can derive the expected phenotypic traits of the ancestors—antimicrobial resistance, metabolic activities, even ecological lifestyle. These inferred traits can then be compared to historical records of Earth evolution or pathogen epidemic spread to try and find causal relations between biological activity and the course of events (David and Alm 2011; Holden et al. 2013), or be considered in support of further ancestral reconstruction, such as scenarios of ecological niche colonisation (Lassalle et al. 2017).

References

Abby SS, Tannier E, Gouy M, Daubin V (2010) Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. BMC Bioinformatics 11:324. https://doi.org/10.1186/1471-2105-11-324

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1000262

Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Angiuoli SV, Salzberg SL (2011) Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 27:334–342

Aziz RK, Bartels D, Best AA et al (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75. https://doi.org/10.1186/1471-2164-9-75

Bankevich A, Nurk S, Antipov D et al (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021

Bérard S, Gallien C, Boussau B et al (2012) Evolution of gene neighborhoods within reconciled phylogenies. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts374

Boussau B, Karlberg EO, Frank AC et al (2004) Computational inference of scenarios for alpha-proteobacterial genome evolution. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.0400975101

Buckee C, Jolley K, Recker M et al (2008) Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. Proc Natl Acad Sci USA 105:15082–15087. https://doi.org/10.1073/pnas.0712019105

Castillo-Ramírez S, Harris SR, Holden MTG et al (2011) The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog 7:e1002129. https://doi.org/10.1371/journal.ppat.1002129

Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205. https://doi.org/10.1038/nrg2526

Collins C, Didelot X (2018) A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol 14: e1005958. https://doi.org/10.1101/140798

Croucher NJ, Page AJ, Connor TR et al (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15. https://doi.org/10.1093/nar/gku1196

Csurös M (2008) Ancestral reconstruction by asymmetric Wagner parsimony over continuous characters and squared parsimony over distributions. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)

Csurös M, Miklós I (2009) Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol 26:2087–2095. https://doi.org/10.1093/molbev/msp123

Darling AE, Miklós I, Ragan MA (2008) Dynamics of genome rearrangement in bacterial populations. PLoS Genet 4:e1000128. https://doi.org/10.1371/Citation

Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147. https://doi.org/10.1371/journal.pone.0011147

David LA, Alm EJ (2011) Rapid evolutionary innovation during an Archaean genetic expansion. Nature. https://doi.org/10.1038/nature09649

Didelot X, Maiden MCJ (2010) Impact of recombination on bacterial evolution. Trends Microbiol 18:315–322. https://doi.org/10.1016/j.tim.2010.04.002

Didelot X, Wilson DJ (2015) ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11:e1004041. https://doi.org/10.1371/journal.pcbi.1004041

Didelot X, Achtman M, Parkhill J et al (2007) A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res 17:61–68. https://doi.org/10.1101/gr.5512906.1

Didelot X, Darling AE, Falush D (2009) Inferring genomic flux in bacteria. Genome Res 19:306–317. https://doi.org/10.1101/gr.082263.108.clearly

Didelot X, Lawson DJ, Darling AE, Falush D (2010) Inference of homologous recombination in bacteria using whole-genome sequences. Genetics 186:1435–1449. https://doi.org/10.1534/genetics.110.120121

Didelot X, Méric G, Falush D, Darling AE (2012) Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics 13:256. https://doi.org/10.1186/1471-2164-13-256

Didelot X, Walker AS, Peto TE et al (2016) Within-host evolution of bacterial pathogens. Nat Rev Microbiol 14:150–162. https://doi.org/10.1038/nrmicro.2015.13

Dingle KE, Elliott B, Robinson E et al (2014) Evolutionary history of the *Clostridium difficile* pathogenicity locus. Genome Biol Evol 6:36–52. https://doi.org/10.1093/gbe/evt204

Donnelly P, Tavare S (1995) Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401–421

Duchemin W, Anselmetti Y, Patterson M et al (2017) DeCoSTAR: reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. Genome Biol Evol. https://doi.org/10.1093/gbe/evx069

Everitt RG, Didelot X, Batty EM et al (2014) Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. Nat Commun 5:3956. https://doi.org/10.1038/ncomms4956

Fisher RA (1931) XVII—the distribution of gene ratios for rare mutations. Proc R Soc Edinburgh. https://doi.org/10.1017/S0370164600044886

Griffiths RC, Marjoram P (1997) An ancestral recombination graph. Prog Popul Genet Hum Evol (Minneapolis, MN, 1994) 87:257–270

Griffiths R, Tavare S (1994) Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc B Biol Sci 344:403–410

Hedge J, Wilson DJ (2016) Practical approaches for detecting selection in microbial genomes. PLoS Comput Biol 12:e1004739. https://doi.org/10.1371/journal.pcbi.1004739

Holden MTG, Hsu L-Y, Kurt K et al (2013) A genomic portrait of the emergence, evolution and global spread of a methicillin resistant *Staphylococcus aureus* pandemic. Genome Res 23:653–664

Huelsenbeck JP, Nielsen R, Bollback JP (2003) Stochastic mapping of morphological characters. Syst Biol. https://doi.org/10.1080/10635150390192780

Jacox E, Chauve C, Szöllősi GJ et al (2016) ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. Bioinformatics 32:2056–2058. https://doi.org/10.1093/bioinformatics/btw105

Jolley KAA, Maiden MCJ (2010) BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595. https://doi.org/10.1186/1471-2105-11-595

Joy JB, Liang RH, Mccloskey RM et al (2016) Ancestral reconstruction. PLoS Comput Biol 12: e1004763. https://doi.org/10.1371/journal.pcbi.1004763

Kingman JFC (1982) The coalescent. Stoch Process Appl 13:235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Kislyuk AO, Haegeman B, Bergman NH, Weitz JS (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics 12:32. https://doi.org/10.1186/1471-2164-12-32

Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. Philos Trans R Soc B Biol Sci 361(1475):1929–1940

Kurtz S, Phillippy A, Delcher AL et al (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12

Lassalle F, Planel R, Penel S et al (2017) Ancestral genome estimation reveals the history of ecological diversification in agrobacterium. Genome Biol Evol 9:3413–3431. https://doi.org/10.1093/gbe/evx255

Lawrence J (1999) Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr Opin Genet Dev 9(6):642–648

Li L, Stoeckert CJJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9):2178–2189. https://doi.org/10.1101/gr.1224503.candidates

Maddison WP (1997) Gene trees in species trees. Syst Biol 46:523–536. https://doi.org/10.1017/CBO9781107415324.004

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature. https://doi.org/10.1038/351652a0

Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol. https://doi.org/10.1186/1471-2148-3-2

Moran PAP (1958) Random processes in genetics. Math Proc Camb Philos Soc 54:60–71

Nordborg M (2001) Coalescent theory. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of statistical genetics. Wiley, Hoboken, NJ

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304. https://doi.org/10.1038/35012500

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res (Camb) 22:201–204. https://doi.org/10.1017/S0016672308009531

Oliveira PH, Touchon M, Cury J, Rocha EPC (2017) The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun. https://doi.org/10.1038/s41467-017-00808-w

Page AJ, Cummins CA, Hunt M et al (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401:877–884. https://doi.org/10.1038/44766

Patterson M, Szöllosi G, Daubin V, Tannier E (2013) Lateral gene transfer, rearrangement, reconciliation. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-14-S15-S4

Pepperell CS, Casto AM, Kitchen A et al (2013) The role of selection in shaping diversity of natural M. tuberculosis populations. PLoS Pathog 9:e1003543. https://doi.org/10.1371/journal.ppat.1003543

Petersen L, Bollback JP, Dimmic M et al (2007) Genes under positive selection in *Escherichia coli*. Genome Res 17:1336–1343. https://doi.org/10.1101/gr.6254707

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3:380–390. https://doi.org/10.1038/nrg795

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Sheppard SK, Didelot X, Jolley KA et al (2013a) Progressive genome-wide introgression in agricultural Campylobacter coli. Mol Ecol 22:1051–1064. https://doi.org/10.1111/mec.12162

Sheppard SK, Didelot X, Meric G et al (2013b) Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci USA 110:11923–11927. https://doi.org/10.5061/dryad.28n35.

Sims GE, Kim S-H (2011) Whole-genome phylogeny of *Escherichia coli*/Shigella group by feature frequency profiles (FFPs). Proc Natl Acad Sci USA 108:8329–8334. https://doi.org/10.1073/pnas.1105168108

Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. https://doi.org/10.1038/nbt.3988

Szollosi GJ, Boussau B, Abby SS et al (2012) Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc Natl Acad Sci 109:17513–17518. https://doi.org/10.1073/pnas.1202997109

Szollosi GJ, Rosikiewicz W, Boussau B et al (2013) Efficient exploration of the space of reconciled gene trees. Syst Biol 62:901–912. https://doi.org/10.1093/sysbio/syt054

Szollosi GJ, Tannier E, Daubin V et al (2015) The inference of gene trees with species trees. Syst Biol 64:e42–e62. https://doi.org/10.1093/sysbio/syu048

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science. https://doi.org/10.1126/science.278.5338.631

Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3:711–721. https://doi.org/10.1038/nrmicro1234

Touchon M, Hoede C, Tenaillon O et al (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 5:e1000344. https://doi.org/10.1371/journal.pgen.1000344

Vallenet D, Labarre L, Rouy Z et al (2006) MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res. https://doi.org/10.1093/nar/gkj406

Vaughan TG, Welch D, Drummond AJ et al (2017) Inferring ancestral recombination graphs from bacterial genomic data. Genetics 205:857–870. https://doi.org/10.1534/genetics.116.193425

Vos M (2011) A species concept for bacteria based on adaptive divergence. Trends Microbiol 19:1–7

Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. ISME J 3:199–208. https://doi.org/10.1038/ismej.2008.93

Vos M, Wolf AB, Jennings SJ, Kowalchuk GA (2013) Micro-scale determinants of bacterial diversity in soil. FEMS Microbiol Rev 37(6):936–954

Ward MJ, Gibbons CL, McAdam PR et al (2014) Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of *Staphylococcus aureus* clonal complex 398. Appl Environ Microbiol 80:7275–7282. https://doi.org/10.1128/AEM.01777-14

Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. Genetics 150:499–510

Wilson DJ, McVean G (2006) Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 172:1411–1425. https://doi.org/10.1534/genetics.105.044917

Wiuf C, Hein J (2000) The coalescent with gene conversion. Genetics 155:451–462

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. Nat Rev Genet 13:303–314. https://doi.org/10.1038/nrg3186

Yang C, Cui Y, Didelot X et al (2018) Why panmictic bacteria are rare. bioRxiv. https://doi.org/10.1101/385336

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. https://doi.org/10.1101/gr.074492.107

Zhou Z, McCann A, Litrup E et al (2013) Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* Serovar Agona. PLoS Genet 9:e1003471. https://doi.org/10.1371/journal.pgen.1003471

Pangenomes and Selection: The Public Goods Hypothesis

James O. McInerney, Fiona J. Whelan, Maria Rosa Domingo-Sananes, Alan McNally, and Mary J. O'Connell

Abstract The evolution and structure of prokaryotic genomes are largely shaped by horizontal gene transfer. This process is so prevalent that DNA can be seen as a public good—a resource that is shared across individuals, populations, and species. The consequence is a network of DNA sharing across prokaryotic life, whose extent is becoming apparent with increased availability of genomic data. Within prokaryotic species, gene gain (via horizontal gene transfer) and gene loss results in pangenomes, the complete set of genes that make up a species. Pangenomes include core genes present in all genomes, and accessory genes whose presence varies across strains. In this chapter, we discuss how we can understand pangenomes from a network perspective under the view of DNA as a public good, how pangenomes are maintained in terms of drift and selection, and how they may differ between prokaryotic groups. We argue that niche adaptation has a major impact on pangenome structure. We also discuss interactions between accessory genes within genomes, and introduce the concepts of 'keystone genes', whose loss leads to concurrent loss of other genes, and 'event horizon genes', whose acquisition may lead to adaptation to novel niches and towards a separate, irreversible evolutionary path.

Keywords Pangenomes · Accessory genes · Epistasis · Public goods

J. O. McInerney (✉) · F. J. Whelan · M. R. Domingo-Sananes · M. J. O'Connell
School of Life Sciences, The University of Nottingham, Nottingham, UK
e-mail: james.mcinerney@nottingham.ac.uk

A. McNally
Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_7

1 Introduction

Horizontal Gene Transfer (HGT) is the most important force affecting evolutionary change in prokaryotes, and its pervasiveness has resulted in a vast global network of connectivity between microorganisms. DNA is available for horizontal acquisition by prokaryotes in a variety of ways: conjugative plasmids (Grohmann et al. 2003; Lederberg and Tatum 1946) facilitate the transfer of DNA directly from cell to cell, phage can facilitate the indirect movement of DNA from one prokaryotic cell to another by generalised transduction (Zinder and Lederberg 1952), and gene transfer agents (GTAs) facilitate gene transfer by cell lysis. In some Archaea, we even see the formation of networks of connections between individuals that can lead to the formation of heterodiploid cells and recombination between the parental cells' genomes (Naor and Gophna 2013). Another important mechanism is direct acquisition of DNA through transformation. Extracellular DNA has a ubiquitous distribution in natural environments from hydrothermal vents, to freshwater, soil, and sediment (Nagler et al. 2018), as well as in the biofilms (Steinberger and Holden 2005) that line our sewage pipes (Vincke et al. 2001), contaminate hospital equipment (Stickler 2008), associate with tooth decay (Potera 1999), and much more. Therefore, DNA can be shared and used among organisms and effectively becomes a public good. All these mechanisms result in a DNA-sharing network that has probably existed since before life evolved to become cellular and will likely remain an important part of prokaryotic biology for as long as there are prokaryotes.

With the advent, and subsequent accessibility, of next-generation sequencing technologies (Shendure et al. 2017), it became apparent that gene presence–absence variability within a species (i.e. strain-to-strain variability) was much larger than expected (Tettelin et al. 2005). For example, when the first three *Escherichia coli* genomes were sequenced, only 39.2% of their protein-coding genes were found to be common to all three genomes (Welch et al. 2002). In a more recent study involving 1524 *Pseudomonas aeruginosa* genomes, only 3% of genes were found to be shared (i.e. 'core') across all strains, with the remaining 97% being variably present in a subset of strains (Karasov et al. 2018). The existence of this variability in gene content within what we regard as single prokaryotic species led to the concept of a pangenome, the complete set of genes that are present in a given species (Tettelin et al. 2005). This set of genes is usually divided into two categories: core genes, that are present across all individuals in a species, and accessory genes, whose presence varies between individuals or strains (Tettelin et al. 2005; Welch et al. 2002; Karasov et al. 2018; Laing et al. 2010). The pangenome concept revolutionises our thinking, since it means considering organisms like *Escherichia coli* not only in terms of the thousand or so genes that are common to all members, but also in terms of the 100,000 or so genes that are found in at least one, but not all, *E. coli* genomes (Land et al. 2015). This new information on the structure of the prokaryotic world has meant that we have to think about 'units' of selection (Okasha 2006) in different ways. In this chapter, we will outline some of the ways in which we can think about pangenomes and what this means for biology. Although our focus is on prokaryotes, it should be noted that some eukaryotes also have pangenomes. For example, a high degree of gene presence–absence polymorphism has been found in different genome sequences of humans (Sherman et al. 2019), cultivated rice (Wang et al. 2018; Hubner et al. 2019), sunflower (Hubner et al. 2019), and in the widespread coccolithophore *Emiliania huxleyi (*Read et al. 2013*)*.

2 Pangenome Properties

As a consequence of the merging of genetic information through HGT and the existence of pangenomes, our thinking about the evolutionary history of prokaryotic genomes has changed. In fact, it is more relevant to think *not* of the evolutionary history of a genome, but rather the evolutionary histories of the various parts of a genome, since these histories can be different (Bapteste et al. 2009). The phylogenetic relationships inferred by a single gene, no matter how important that gene, rarely reflects the evolutionary history of the suite of organisms under consideration. This idea was codified by Darwin in 'The Origin' when he said: '*The importance, for classification, of trifling characters, mainly depends on their being correlated with several other characters of more or less importance*' (Darwin 1860). In other words, the notion of homoplastic characters (i.e. characters whose similarity is due to convergent evolution) is an old idea and characters can differ in what they suggest is the proper classification of an organism. Though Darwin did not know about DNA or HGT, the warning about character congruence and classification still holds true today and perhaps even more so because of HGT and the non-tree likeness of this process.

The pangenomes of different prokaryotic groups differ. Transformation, transduction, and conjugation contribute to shuffling variably sized portions of genomes through both homologous and non-homologous recombination. The frequency of the different mechanisms likely depends on environmental conditions, lifestyle, and cell biology (i.e. the molecular mechanisms present in particular cells or taxa) (Hanage 2016). Therefore, under different conditions, HGT and recombination can in principle range from non-existent to widespread, resulting in primarily clonal or panmictic groups, respectively (Yang et al. 2019). Furthermore, recombination barriers, both within and between species, can be fuzzy and potentially differ for different parts of the genome. This can make the delineation of populations or of species more complicated in prokaryotes, when compared to animals, for example (Hanage 2013). However, it has been suggested that natural species boundaries do exist in prokaryotes and that they can be defined (Bobay and Ochman 2017). On the whole, HGT and DNA recombination in prokaryotes can have similar consequences to sexual reproduction in eukaryotes: removing deleterious mutations, thereby avoiding Muller's ratchet or mutational meltdown, while also offering a mechanism for bringing together advantageous mutations in different genes or parts of the genome. But crucially in prokaryotes, recombination can both remove and add a hugely variable number of genes to a genome, thereby affecting the overall gene repertoire rather than simply modifying existing genes by point mutation. That is,

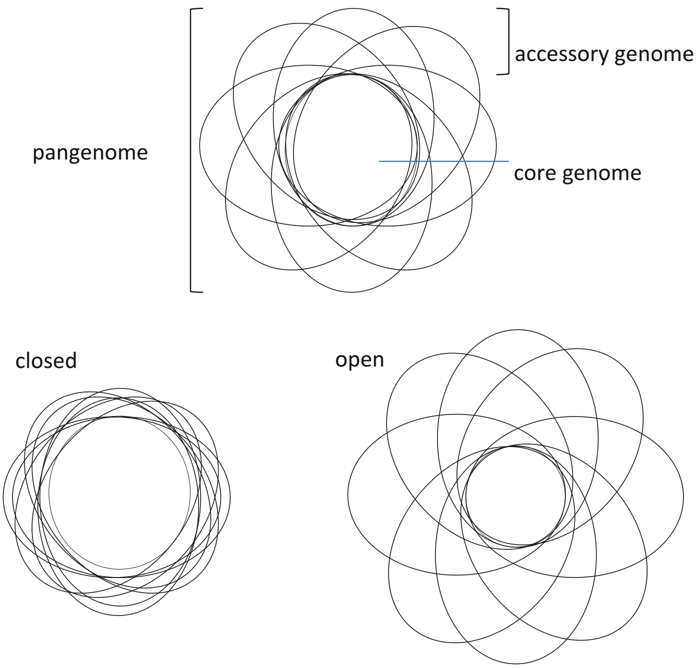


Fig. 1 An illustration of how the rate at which new accessory genes are discovered as increasing numbers of genomes are sequenced. For species with open pangenomes, the rate of accessory gene discovery continually increases, while for closed pangenomes, this rate plateaus quickly

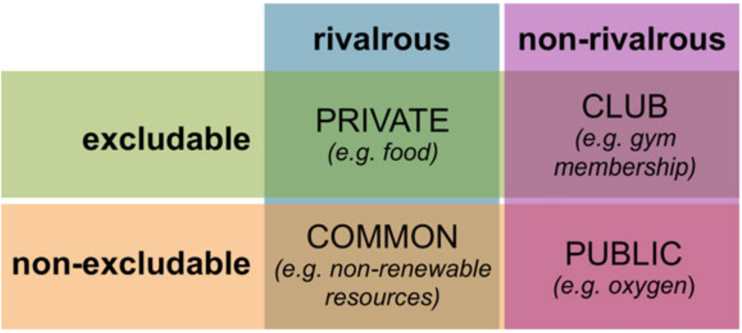
recombination in prokaryotes often results in insertions or deletions, while in eukaryotes it tends to swap alleles between chromosomes.

Pangenomes differ in the degree to which they are 'open' or 'closed'. Species that share almost all genes with each other (i.e. have very little strain-to-strain gene content dissimilarity) having a large 'core' and small 'accessory' genome, are considered to have closed pangenomes (McInerney et al. 2017). In contrast, species can have open pangenomes in which gene content varies appreciably from one genome to another (McInerney et al. 2017) (see Fig. 1). Though we know the openness of prokaryotic pangenomes varies greatly from one species to the next (Tettelin et al. 2005), our estimates of openness can be affected by the available genomic data (i.e. the number of accessory genes is expected to increase as more strain information becomes available). As such, openness can be measured by modelling the number of accessory genes as a function of the number of sequenced genomes (Tettelin et al. 2005) (see also Chap. 1). The first analysis of openness found that eight *Streptococcus agalactiae* genomes were not enough to uncover all possible accessory genes and predicted that new genes would be found with every additional genome, leading to an essentially infinite pangenome. In contrast, the number of new accessory genes in *Bacillus anthracis* dropped to zero after the incorporation of only four genomes to the study of its pangenome (Tettelin et al. 2005). Therefore, accurate measurements of pangenome openness depend on sampling the broad diversity of genomes in a given species, and such measurements should ideally account for core genome diversity and the phylogenetic relationships between those genomes.

3 Public Goods

The idea that DNA functions as a public good (Erwin 2015; McInerney et al. 2011a; McInerney and Erwin 2017) stems from the fact that HGT makes DNA available to other 'users' and this process has structured a great deal of the life on this planet, both cellular and viral (Bapteste et al. 2012, 2013). Integration of a new DNA sequence into a genome can only be successful if the source organism and the recipient organism can both make use of this DNA in some way. Carl Woese referred to the universal genetic code as being the 'lingua franca' of genetic commerce (Woese 2002). HGT has been observed in almost all known phyla, though HGT seems to be reduced in frequency among eukaryotes and perhaps reduced further in multicellular organisms (Schonknecht et al. 2014; McInerney et al. 2014; Ku et al. 2015). As a consequence of HGT, there is no universal Tree of Life, and instead there is a network of life reflecting the vertical and horizontal movements of genetic information (Bapteste et al. 2012, 2013; Corel et al. 2018).

Our current appreciation of evolutionary history in prokaryotes and the observations of pangenomes has led us to consider what metaphors might be appropriate for representing, modelling, and understanding life on the planet. A variety of alternatives to the tree metaphor, such as 'cobwebs of life' (Ge et al. 2005) or 'rhizome of life' (Merhej et al. 2011), have been used. However, some of us have proposed to depart from a way of thinking that inherently depends on a particular kind of diagram. Instead we have advocated a focus on the fundamental process of HGT, and the fact that it constructs new genomes in the same way that, say, a furniture manufacturing plant might bring together different materials in order to construct a new kind of chair, or in the way that a football team might substitute one player for another. As mentioned above, Woese suggested that HGT could be thought of in commercial terms (Woese 2002), and a logical extension to this line of thinking is that DNA acts as though it is a 'public good' (McInerney et al. 2011a, b; McInerney and Erwin 2017). Briefly, in the theory of goods, Nobel laureate Paul Samuelson initially described two kinds of goods thus: '[. . .] I explicitly assume two categories of goods: ordinary private consumption goods which can be parcelled out among different individuals [. . .] and collective consumption goods [. . .] which all enjoy in common in the sense that each individual's consumption of such a good leads to no subtraction from any other individual's consumption of that good [. . .]' (Samuelson 1954). Since then, the concept has been expanded so that four kinds of goods are recognised—private goods, public goods, club goods, and common goods (McInerney et al. 2011a), based on whether goods are rivalrous and/or excludable. The criteria for each of the classifications are contained in Fig. 2, along with examples of goods that fall easily into each of these categories. A 'good' is said to be rivalrous if its consumption by one consumer prevents simultaneous consumption by other consumers, and a 'good' is said to be excludable if it is possible to prevent



	rivalrous	non-rivalrous
excludable	PRIVATE *(e.g. food)*	CLUB *(e.g. gym membership)*
non-excludable	COMMON *(e.g. non-renewable resources)*	PUBLIC *(e.g. oxygen)*

Fig. 2 The nature of Goods. Goods can fall into four different categories—private, club, common, and public according to whether they are rivalrous or non-rivalrous, and excludable or non-excludable. The figure also gives some examples of goods that easily fall into each of these four categories

others from having access to it. DNA possesses the property of being non-excludable (e.g. the DNA of any individual is made available, at least at the time of death of the cell or the individual) and it is also non-rivalrous in a practical sense, given that the amount of DNA that is produced by any given species cannot realistically be used up by any consumer. This perspective is useful in the sense that viewing genome evolution as a process of building functioning tools (i.e. new kinds of organisms) allows us to ask questions that would not make much sense if we used 'tree-thinking' (Bapteste et al. 2013; Dagan and Martin 2009). Tree-thinking inherently supposes that genes came to be in a genome because all the genes have been inherited through the same lineage of descent—a process that infers that genes are 'private' to a clade. 'Goods-thinking', on the other hand, frees us to think more about why the particular set of genes that we observe in a genome are there, rather than some other set of genes. We do not assume that any gene is a private good, exclusively found in a particular species or clade, with other organisms excluded from accessing the segment of DNA. Goods-thinking infers that a genome has evolved to be the way it is through vertical inheritance from a common ancestor, but also through the horizontal acquisition of genes, with the rate of gain (and loss) of genes being modified by the influences of random drift, selection, and demography. Goods-thinking, therefore, needs some new tools, outside of the framework of the bifurcating phylogenetic tree, in order to properly analyse gene and genome evolution (Bapteste et al. 2009). Here we deal specifically with the pangenome's part of Goods Thinking theory.

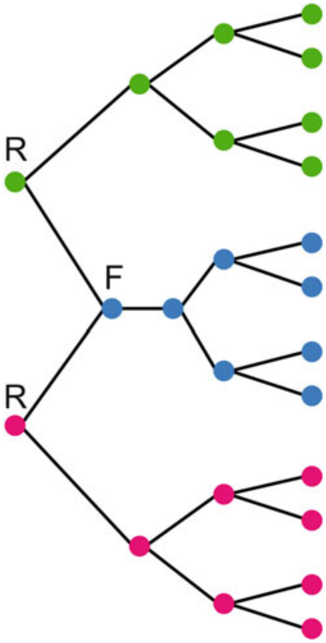
4 Analyses of Pangenomes

Because of the fluidity of genomes, caused by accessory gene gain and loss, the analysis of pangenomes lends itself more suitably to networks than to phylogenies. Networks are mathematical graphs represented by nodes, or vertices, which are connected by edges, or lines, if-and-only-if a relationship exists between them. Networks are widely used in ecology—and in biology in general—to represent, for example food webs (Dunne et al. 2002), social interactions (Robins et al. 2007), nutrient/energy flows (Allesina et al. 2005), and cooperation between members in a population (Jain and Krishna 2001). Networks can have edges that are either directed (often shown as an arrow) or undirected, depending on whether the relationship that connects the nodes has directionality (e.g. to connect an organism to their food source in a food web). The study of networks, or graphs (i.e. graph theory) dates to at least 1735 (Skiena 2008; Compeau et al. 2011) and has advanced rapidly due to its applications in computer science, engineering, physics, and biology. The public goods nature of DNA makes a network structure ideal to uncover patterns and processes of evolution in ways where phylogenetic trees would be somewhat lacking, since phylogenetic trees do not infer lateral movement of genetic material. The analysis of features contained within the graphs such as non-transitive triplets, or nodes with identical incident edges can reveal patterns of recombination or gene sharing (Bapteste et al. 2012; Corel et al. 2018; Meheust et al. 2018).

In the analysis of pangenomes, networks are often *k-partite* or *multi-partite*, meaning that their nodes can be coloured using *k* colours such that no node is directly connected to another with the same colour (Pavlopoulos et al. 2018). A special case of *k-partite* graphs is *bipartite* or two colourable graphs. In pangenome analyses, bipartite graphs usually connect genomes to their constituent genes (Corel et al. 2018). Bipartite networks have been used previously to identify the levels of gene sharing within microbial genomes (Corel et al. 2018), to characterise the capacity of accessory genes in metabolic networks (Goyal 2018), and to interrogate gene presence/absence patterns and coincident relationships (McNally et al. 2016).

Especially relevant for genome evolution is the N-rooted fusion graph (Haggerty et al. 2014). This graph differs from a phylogenetic tree due to the presence of more than one root node (a node that depicts the point-of-origin of all operational taxonomic units in the graph) and the presence of at least one internal node in the graph where the in-degree of the node (the number of edges pointing towards that node) is greater than 1 and the out-degree of the node (the number of edges emerging from that node) is 1 (Fig. 3). In other words, the merging of genetic material inherently means that the graph needs more than a single origin or root. It also means that the point at which the material merged must be represented by a merger, or fusion node (Fig. 3). The various components of the internal structure of an N-rooted fusion graph can be determined by the usual phylogenetic approaches [i.e. parsimony, likelihood, or distance matrix methods (Felsenstein 2003)]. The complete N-rooted graph is then constructed by merging of these individual phylogenetic trees, by constructing fusion nodes at the appropriate places (Haggerty et al. 2014).

Fig. 3 An N-Rooted Fusion Graph. This kind of branching diagram can be used to illustrate the merging of evolving objects. The nodes labelled R indicate the root nodes for this graph. Each root node depicts the root for a different kind of gene. The node labelled F indicated the fusion node. The different node colours indicate different gene families, with the blue nodes indicating that they are a fusion family



5 How Are Pangenomes Maintained?

Because acquired DNA can function across multiple organisms—facilitating it to become a public good—HGT into some individuals in a population creates diversity within that species. Transferred sequences will be present in a subset of the population's genomes and absent in the rest (McNally et al. 2016), becoming raw material for natural selection (see Fig. 4). Multiple iterations of this process have most likely resulted in the observed pattern of hugely varying gene content across conspecific genomes (Welch et al. 2002; Lukjancenko et al. 2010; Koonin and Wolf 2008). Maintenance of the observed high levels of variation requires an explanation, because, while we know that transformation, conjugation, and transduction introduce this presence–absence variation, it is expected that both natural selection and genetic drift would remove this kind of genetic variation from populations. In terms of sequence variation within populations, different mechanisms have been proposed to explain the maintenance of diversity. These mechanisms range from relatively trivial explanations, such as the existence of a balance between the rates at which new variants arise in populations (by mutation, for example) and the rates at which they are removed, to more exotic mechanisms such as heterozygote advantage, interactions between genotypes and different environments, and negative frequency-dependent selection (Hahn 2018). Although most of these explanations have been developed in order to account for high levels of genetic diversity in diploid, sexually reproducing eukaryotes, some of these mechanisms can also help

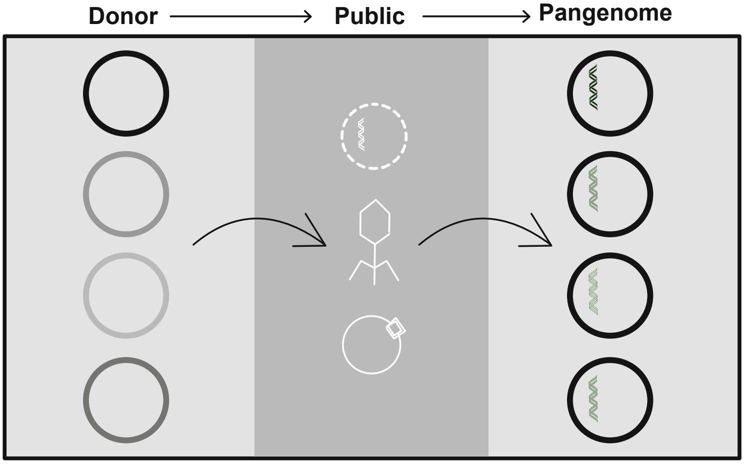


Fig. 4 Prokaryotic DNA becomes a public good upon cell death or when the DNA is taken from the cell via phage or plasmids. Pangenomes can then accrue via the differential acquisition of these public goods

us to understand genetic variation in prokaryotes. However, understanding the existence and maintenance of pangenomes has its own particular challenges.

A key element to be considered when we speak about mechanisms that maintain variability in gene content in prokaryotic populations is the fitness effect that these accessory genes have on individuals. We will likely find examples of particular genes whose presence is neutral, deleterious, or adaptive in most genomes; we are already familiar with genes in the latter class such as those conferring antibiotic resistance and pathogenicity islands (Sheppard et al. 2018). However, an interesting question to think about is whether accessory genes on average contribute to fitness (or under which circumstances they may be adaptive), and which mechanisms have led to their patchy occurrence in genomes. Depending on the average fitness effect of accessory genes, different mechanisms could be governing their presence.

If accessory genes are mostly deleterious, which could be the case if they are predominantly selfish or parasitic, then the patchy presence patterns that we observe could reflect a constant arms race between these selfish elements and the host genome (somewhat equivalent to the Red Queen hypothesis for maintenance of variability in populations of interacting hosts and pathogens). Although this pattern may be responsible for a proportion of accessory genes, it is very unlikely that this explains most of the observed variability and the existence of pangenomes, partly because many accessory genes are not related to selfish elements and appear to be involved in multiple cellular functions (McNally et al. 2016; Sheppard et al. 2018).

If accessory genes are usually neutral in terms of fitness, eventually they would be randomly fixed or lost in different populations due to genetic drift, particularly if recombination is rare. A neutralist model for pangenomes implies that we see presence–absence variation because there is a random ‘rain’ of genes constantly being acquired and we observe their presence in a genome because they have either not had enough time to drift to fixation or to be lost again. This kind of model implies that neither the gain nor the loss of accessory genes has a fitness effect (Baumdicker et al. 2012), a situation that seems contradicted by the observation of both prophage (Nanda et al. 2015) and antibiotic resistance (Her and Wu 2018) genes affecting fitness. A recent study (Andreani et al. 2017) showed a correlation between pangenome fluidity and synonymous variation, which was taken to imply that genome content diversity is mostly neutral. The implication was that synonymous diversity arises in the absence of selection and if this correlates with genome fluidity, then genome fluidity is also neutral. The problem with this model is that synonymous diversity in prokaryotes is not necessarily neutral, and we see stronger selection on synonymous codon usage in organisms with large effective population size (N_e) (Sharp et al. 1993), so the correlation between large N_e and genome fluidity is unlikely to be a consequence of drift alone.

Recently, a drift-barrier model for pangenome evolution has been proposed (Bobay and Ochman 2018). The authors observed a positive correlation between pangenome size and N_e (using two independent measures of N_e for different bacterial species). In contrast to Andreani et al. (2017) they propose that, on average, accessory genes make a positive contribution to fitness. Based on nearly neutral evolutionary theory, they then explain the correlation between N_e and pangenome size by the loss of slightly advantageous genes in populations with small N_e. Therefore, populations with large N_e would maintain a larger number of accessory genes. However, while this may help explain larger genome size (i.e. the maintenance of more genes), it does not necessarily explain diversity in gene content in different individuals from the same population, since those slightly advantageous genes would be expected to eventually fix in the population. Furthermore, the authors did not deal with the likelihood that, on occasion, these advantageous genes would result in sweeps to fixation. The problem with this model is outlined in simulations by Niehus et al. (2015). As some of us have previously proposed (McInerney et al. 2017), some of the basics of this drift-barrier model, if combined with niche adaptation, can go further in explaining the maintenance of genome content diversity. Under the adaptive pangenomes model of McInerney et al. (2017), accessory genes make, on average, a positive contribution to fitness, and this contribution may be niche dependent. Therefore, genes are maintained in the niches where they are beneficial and lost in others. However, ongoing migration would still allow recombination in other parts of the genome, and thus maintenance of large N_e, at least for the core genome.

In line with the McInerney et al. (2017) model of pangenome maintenance by a combination of drift and niche-dependence, there is evidence that at least a significant fraction of accessory genes are beneficial and involved in niche adaptation (Bruns et al. 2018; Rubino et al. 2017; McInerney 2013). The adaptability of prokaryotes means that they occupy niches all over the planet—including oceans (Sunagawa et al. 2015), ice sheets (Anesio et al. 2017), and salt flats (Caton et al. 2004), as well as ecosystems deep within the earth's crust (Chivian et al. 2008), and on and within our own bodies (The Human Microbiome Project Consortium 2012).

Some ‘specialist’ prokaryotic species focus on one, specific niche; for example *Buchnera aphidicola* is an endosymbiont that forms an obligate association with aphids (van Ham et al. 2003). Such specialists would likely have little to gain from extensive gene content diversity, possibly explaining the relative closeness of some species pangenomes. For example, *Tropheryma whipplei*, an intracellular human pathogen and the causative agent of Whipple’s disease (Gorvel et al. 2010), has an extremely restricted pangenome (Fenollar et al. 2014), and smaller than average N_e (Bobay and Ochman 2018). In contrast, ‘generalist’ prokaryotic species can occupy many of the niches made available to them. *Escherichia coli* has been identified in several different kinds of environments including the gut and urinary tract of humans, and indeed other warm- and cold-blooded animals (Tenaillon et al. 2010), as well as soil, sediment, and water (Savageau 1983). In order to occupy such variable environments, these species must be able to adapt to different carbon and nitrogen sources (Bertin et al. 2011), to evade various antibiotic pressures (Sáenz et al. 2004), and to utilise different types of respiration depending on oxygen availability (Jones et al. 2007). Recent work on the metabolic potential of accessory genes has identified a correlation between the number of novel metabolites that a given strain can synthesise and the openness of their pangenome, suggesting that the acquisition of such genes is adaptive (Goyal 2018). Other scenarios where variation in accessory genes is actively maintained by selection include negative frequency-dependent selection (Corander et al. 2017) where a major allele (gene presence or absence in our case) is at a disadvantage compared with the minor allele (the other character state). For example, in the case of vaccine programmes, it is likely that a vaccine targeting a non-essential accessory gene will confer a selective advantage on strains that do not have that accessory gene (Azarian et al. 2018). Bacteriophages may have a similar effect on non-essential attachment proteins and other cellular components. Finally, it is also the case that a particular gene may be beneficial in a specific niche when another gene is present, but not so when that partner is absent. This co-dependency of genes for fitness/adaptation to a particular niche will manifest particular patterns of co-occurrence in genomes (Cohen et al. 2013).

Notwithstanding the argument being made here that pangenomes are, on average, constructed and maintained by niche adaptation, we are still a long way from having enough data to say that this understanding is true in all cases. To assess whether neutralist or selectionist scenarios warrant greater or lesser support in different prokaryotic species and populations, we need more genomic data and information on population structure, levels of migration and recombination, and the distribution of fitness effects of accessory genes in different niches or environments. This requires deep sampling of prokaryotic genomes across space (within and between niches) and ideally along time. Recording of information on as many environmental variables as possible would also be highly advantageous for understanding which factors influence the evolution of pangenomes.

6 Keystone Genes and Event Horizon Genes

The dynamics of accessory gene repertoires is clearly a subject of great interest in microbiology. We have a poor understanding of how these repertoires are structured and what influences their content, how they grow and are maintained. The process of gene loss is also poorly understood. We have outstanding questions about what we might term 'keystone genes', those genes that play a central role in determining what other genes might be successful in a genome. This keystone gene concept is analogous to the keystone species concept in macroecology (Paine 1969); keystone species are those whose presence or absence can result in a major shift in the make-up of a particular ecosystem, often resulting in ecosystem collapse, if the keystone species leaves or goes extinct (Estes et al. 1978).

In a related, but slightly different context, we might consider the case of 'event horizon' genes. To give an example of the possible existence of such genes, we can consider the evolution by gene acquisition of Archaeal halophiles from an ancestor that was a methanogen (Nelson-Sathi et al. 2012). This transition must have involved the rapid acquisition of a large number of genes. Whereas Haloarchaea are heterotrophic, facultatively anaerobic or aerobic organisms with a phototrophic capability, their ancestors the methanogens are obligately anaerobic, methane-producing, chemolithotrophic archaea. The differences between these two closely related groups illustrate that seismic changes in genome content can occur, but also that the absence of intermediate forms suggests that such changes can come about with great rapidity. This leads us to the question of which genes, when acquired, led to the establishment of the halophile phenotype. In an analogy with astrophysics, we can speculate whether there has been an 'event horizon' or a point of no return, where the acquisition of a particular gene or set of genes permanently converted a methanogen to a halophile. We might imagine that the combination of importers of organic compounds and genes for heterotrophic metabolism marked the point of no return. Indeed, there seems to have been in this case no return, since all halophilic archaea are monophyletic and none have abandoned this lifestyle. Therefore, the order of gene acquisition and gene loss is an important question. Future work will help understand whether these keystone and event horizon genes are common in accessory gene repertoires.

7 Some Conclusions and Future Directions

While evolution has no particular direction, the likely success of a particular genomic sequence relates to the notion of 'unity of purpose'. In this sense, the various components of a biochemical pathway can be said to have unity of purpose—collectively they enable the biological transformation of some important molecules. The components of the translation apparatus similarly have a unity of purpose. As a corollary, we could say that inserting genes that can enable methanogenesis into the same genome as genes that are responsible for importing sugars would not likely lead to a genome with a particularly united purpose—one part of the genome would be dedicated to producing energy by chemolithotrophy, while another part of the genome would be dedicated to a heterotrophic lifestyle. Yet situations like this must surely arise from time to time, given the pervasiveness of HGT. Two great unknowns right now include how often such conflicts arise in nature, and how compatible are the genes we see in genomes. We know that they are compatible enough to give rise to functioning organisms, but we do not know how each individual gene contributes to fitness. Background selection and hitch-hiking Hill-Robertson effects (Hill and Robertson 1966) are mechanisms that can limit the 'impact' of natural selection and allow maintenance of slightly deleterious variants (Price and Arkin 2015), including, we would suppose, accessory genes that have a slightly deleterious fitness effect.

The focus on pangenomes is usually centred on protein-coding genes, but there are several other levels at which pangenomes provide food for thought. An analysis of *E. coli* genomes has revealed that selection on non-coding regions has been instrumental in shaping the success of a particular sequence type (ST131) of the species (McNally et al. 2016). This brings into focus the combinatorial nature of genome structure—that the presence or absence of particular kinds of protein-coding genes, or even RNA-coding genes is only part of the story, and that the 'regulatory pangenome' will be one of the most important future challenges.

Acknowledgements We would like to acknowledge the financial support of the Biotechnology and biological sciences research council (BBSRC) grant number BB/N018044/1. JOM and MOC would like to acknowledge a start-up grant from the School of Life Sciences, University of Nottingham. FJW would like to acknowledge that this project received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 793818.

References

Allesina S, Bodini A, Bondavalli C (2005) Ecological subsystems via graph theory: the role of strongly connected components. Oikos 110(1):164–176

Andreani NA, Hesse E, Vos M (2017) Prokaryote genome fluidity is dependent on effective population size. ISME J 11(7):1719–1721

Anesio AM et al (2017) The microbiome of glaciers and ice sheets. NPJ Biofilms Microbiomes 3:10

Azarian T et al (2018) Prediction of post-vaccine population structure of *Streptococcus pneumoniae* using accessory gene frequencies. bioRxiv. https://doi.org/10.1101/420315

Bapteste E et al (2009) Prokaryotic evolution and the tree of life are two different things. Biol Direct 4(1):34

Bapteste E et al (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. Proc Natl Acad Sci USA 109(45):18266–18272

Bapteste E et al (2013) Networks: expanding evolutionary thinking. Trends Genet 29(8):439–441

Baumdicker F, Hess WR, Pfaffelhuber P (2012) The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol 4(4):443–456

Bertin Y et al (2011) Enterohaemorrhagic *Escherichia coli* gains a competitive advantage by using ethanolamine as a nitrogen source in the bovine intestinal content. Environ Microbiol 13 (2):365–377

Bobay LM, Ochman H (2017) Biological species are universal across life's domains. Genome Biol Evol 9(3):491–501

Bobay L-M, Ochman H (2018) Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol 18(1):153

Bruns H et al (2018) Function-related replacement of bacterial siderophore pathways. ISME J 12 (2):320–329

Caton TM et al (2004) Halotolerant aerobic heterotrophic bacteria from the Great Salt Plains of Oklahoma. Microb Ecol 48(4):449–462

Chivian D et al (2008) Environmental genomics reveals a single-species ecosystem deep within Earth. Science 322(5899):275–278

Cohen O et al (2013) CoPAP: coevolution of presence-absence patterns. Nucleic Acids Res 41(Web Server issue):W232–W237

Compeau PEC, Pevzner PA, Tesler G (2011) Why are de Bruijn graphs useful for genome assembly? Nat Biotechnol 29(11):987

Corander J et al (2017) Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nat Ecol Evol 1(12):1950–1960

Corel E et al (2018) Bipartite network analysis of gene sharings in the microbial world. Mol Biol Evol 35(4):899–913

Dagan T, Martin W (2009) Getting a better picture of microbial evolution en route to a network of genomes. Philos Trans R Soc Lond Ser B Biol Sci 364(1527):2187–2196

Darwin C (1860) On the origin of species by means of natural selection, 2nd edn. John Murray, London

Dunne JA, Williams RJ, Martinez ND (2002) Food-web structure and network theory: the role of connectance and size. Proc Natl Acad Sci USA 99(20):12917–12922

Erwin DH (2015) A public goods approach to major evolutionary transitions. Geobiology 13:308–315

Estes JA, Smith NS, Palmisano JF (1978) Sea otter predation and community organization in Western Aleutian Islands, Alaska. Ecology 59(4):822–833

Felsenstein J (2003) Inferring phylogenies. Oxford University Press, Oxford, p 580

Fenollar F et al (2014) Tropheryma whipplei and Whipple's disease. J Infect 69(2):103–112. https://doi.org/10.1016/j.jinf.2014.05.008

Ge F, Wang LS, Kim J (2005) The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol 3(10):e316

Gorvel L et al (2010) Tropheryma whipplei, the Whipple's disease bacillus, induces macrophage apoptosis through the extrinsic pathway. Cell Death Dis 1(4):e34–e34

Goyal A (2018) Metabolic adaptations underlying genome flexibility in prokaryotes. PLoS Genet 14(10):e1007763

Grohmann E, Muth G, Espinosa M (2003) Conjugative plasmid transfer in gram-positive bacteria. Microbiol Mol Biol Rev 67(2):277–301

Haggerty LS et al (2014) A pluralistic account of homology: adapting the models to the data. Mol Biol Evol 31(3):501–516

Hahn MW (2018) Molecular population genetics. Oxford University Press, Oxford

Hanage WP (2013) Fuzzy species revisited. BMC Biol 11:41

Hanage WP (2016) Not so simple after all: bacteria, their population genetics, and recombination. Cold Spring Harb Perspect Biol 8(7). https://doi.org/10.1101/cshperspect.a018069

Her HL, Wu YW (2018) A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains. Bioinformatics 34(13):i89–i95

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res 8 (3):269–294

Hubner S et al (2019) Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat Plants 5(1):54–62

Jain S, Krishna S (2001) A model for the emergence of cooperation, interdependence, and structure in evolving networks. Proc Natl Acad Sci USA 98(2):543–547

Jones SA et al (2007) Respiration of *Escherichia coli* in the mouse intestine. Infect Immun 75 (10):4891–4899

Karasov TL et al (2018) *Arabidopsis thaliana* and *Pseudomonas pathogens* exhibit stable associations over evolutionary timescales. Cell Host Microbe 24(1):168–179.e4

Koonin EV, Wolf YI (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res 36(21):6688–6719

Ku C et al (2015) Endosymbiotic origin and differential loss of eukaryotic genes. Nature 524 (7566):427–432

Laing C 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

Land M et al (2015) Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15(2):141–161

Lederberg J, Tatum EL (1946) Gene recombination in *Escherichia coli*. Nature 158(4016):558

Lukjancenko O, Wassenaar TM, Ussery DW (2010) Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol 60(4):708–720

McInerney JO (2013) More than tree dimensions: inter-lineage evolution's ecological importance. Trends Ecol Evol 28(11):624–625

McInerney JO, Erwin DH (2017) The role of public goods in planetary evolution. Philos Trans A Math Phys Eng Sci 375(2109). https://doi.org/10.1098/rsta.2016.0359

McInerney JO et al (2011a) The public goods hypothesis for the evolution of life on earth. Biol Direct 6:41

McInerney J, Cummins C, Haggerty L (2011b) Goods-thinking vs. tree-thinking: finding a place for mobile genetic elements. Mob Genet Elem 1(4):1–4

McInerney JO, O'Connell MJ, Pisani D (2014) The hybrid nature of the Eukaryota and a consilient view of life on Earth. Nat Rev Microbiol 12(6):449–455

McInerney JO, McNally A, O'Connell MJ (2017) Why prokaryotes have pangenomes. Nat Microbiol 2:17040

McNally A 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(9):e1006280

Meheust R et al (2018) Formation of chimeric genes with essential functions at the origin of eukaryotes. BMC Biol 16(1):30

Merhej V et al (2011) The rhizome of life: the sympatric *Rickettsia felis* paradigm demonstrates the random transfer of DNA sequences. Mol Biol Evol 28(11):3213–3223

Nagler M et al (2018) Extracellular DNA in natural environments: features, relevance and applications. Appl Microbiol Biotechnol 102(15):6343

Nanda AM, Thormann K, Frunzke J (2015) Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. J Bacteriol 197(3):410–419

Naor A, Gophna U (2013) Cell fusion and hybrids in Archaea: prospects for genome shuffling and accelerated strain development for biotechnology. Bioengineered 4(3):126–129

Nelson-Sathi S et al (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc Natl Acad Sci USA 109(50):20537–20542

Niehus R et al (2015) Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nat Commun 6:8924

Okasha S (2006) Evolution and the levels of selection. Oxford University Press, Oxford

Paine RT (1969) A note on trophic complexity and community stability. Am Nat 103(929):91–93

Pavlopoulos GA et al (2018) Bipartite graphs in systems biology and medicine: a survey of methods and applications. Gigascience 7(4):1–31. https://doi.org/10.1093/gigascience/giy014

Potera C (1999) Forging a link between biofilms and disease. Science 283(5409):1837–1839

Price MN, Arkin AP (2015) Weakly deleterious mutations and low rates of recombination limit the impact of natural selection on bacterial genomes. MBio 6(6):e01302–e01315

Read BA et al (2013) Pan genome of the phytoplankton *Emiliania* underpins its global distribution. Nature 499(7457):209–213

Robins G et al (2007) An introduction to exponential random graph (p∗) models for social networks. Soc Netw 29(2):173–191

Rubino F et al (2017) Divergent functional isoforms drive niche specialisation for nutrient acquisition and use in rumen microbiome. ISME J 11(4):932–944. https://doi.org/10.1038/ismej.2016.172

Sáenz Y et al (2004) Mechanisms of resistance in multiple-antibiotic-resistant *Escherichia coli* strains of human, animal, and food origins. Antimicrob Agents Chemother 48(10):3996–4001

Samuelson PA (1954) The pure theory of public expenditure. Rev Econ Stat 36(4):387–389

Savageau MA (1983) *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. Am Nat 122(6):732–744

Schonknecht G, Weber AP, Lercher MJ (2014) Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. Bioessays 36(1):9–20

Sharp PM et al (1993) Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans 21(4):835–841

Shendure J et al (2017) DNA sequencing at 40: past, present and future. Nature 550(7676):345–353

Sheppard SK, Guttman DS, Fitzgerald JR (2018) Population genomics of bacterial host adaptation. Nat Rev Genet 19(9):549–565

Sherman RM et al (2019) Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet 51(1):30–35

Skiena SS (2008) The algorithm design manual. Springer, London

Steinberger RE, Holden PA (2005) Extracellular DNA in single- and multiple-species unsaturated biofilms. Appl Environ Microbiol 71(9):5404–5410

Stickler DJ (2008) Bacterial biofilms in patients with indwelling urinary catheters. Nat Clin Pract Urol 5(11):598–608

Sunagawa S et al (2015) Structure and function of the global ocean microbiome. Science 348 (6237):1261359

Tenaillon O et al (2010) The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol 8(3):207–217

Tettelin H et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. Proc Natl Acad Sci USA 102 (39):13950–13955

The Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486(7402):207–214

van Ham RCHJ et al (2003) Reductive genome evolution in *Buchnera aphidicola*. Proc Natl Acad Sci USA 100(2):581–586

Vincke E, Boon N, Verstraete W (2001) Analysis of the microbial communities on corroded concrete sewer pipes—a case study. Appl Microbiol Biotechnol 57(5–6):776–785

Wang W et al (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557(7703):43–49

Welch RA et al (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA 99(26):17020–17024

Woese CR (2002) On the evolution of cells. Proc Natl Acad Sci 99(13):8742–8747

Yang C et al (2019) Why panmictic bacteria are rare. bioRxiv. https://doi.org/10.1101/385336

Zinder ND, Lederberg J (1952) Genetic exchange in *Salmonella*. J Bacteriol 64(5):679–699

A Pangenomic Perspective on the Emergence, Maintenance, and Predictability of Antibiotic Resistance

Stephen Wood, Karen Zhu, Defne Surujon, Federico Rosconi, Juan C. Ortiz-Marquez, and Tim van Opijnen

Abstract The rapidly expanding number of sequenced bacterial strains and species, and the ongoing curation of bacterial pangenomes has uncovered unexpected complexities in understanding and addressing antibiotic resistance in the context of the pangenome. It is becoming apparent that differences in the genetic background can cause species and strain-specific responses to the same antibiotic, triggering differential selective pressures and thereby strain or species-specific adaptive outcomes. In this chapter, we consider how the pangenome, on a between and within species level, can affect the response to antibiotics and the development of resistance as well as the role selective pressures such as antibiotics play in shaping and maintaining the pangenome. We review the tools that are used to study antibiotic resistance within a pangenomic context, highlight recent findings, discuss strategies for predicting the emergence of resistance and consider how effective therapies can be developed in the context of the pangenome.

Keywords Pangenome · Antibiotic resistance · Genomics · High-throughput tools · Adaptive evolution · Network analyses · Epistasis · Predictions · Machine learning

1 Introduction

Antibiotic resistance is a naturally occurring phenomenon that can be found in environments containing antibiotic-producing microorganisms, even in the absence of human activity (D'Costa et al. 2006). While antibiotic resistance is rampant in livestock and is a confounding factor in the emergence and spread of resistance into the human population, most research focuses on bacterial pathogens affecting humans and in particular the ESKAPE pathogens (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas*

S. Wood · K. Zhu · D. Surujon · F. Rosconi · J. C. Ortiz-Marquez · T. van Opijnen (✉)
Biology Department, Boston College, Chestnut Hill, MA, USA
e-mail: vanopijn@bc.edu

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_8

aeruginosa, and *Enterobacter* species) as well as *Mycobacterium tuberculosis* (Santajit and Indrawattana 2016). Instrumental in the development of resistance is a bacterium's inherent ability to survive exposure to low antibiotic concentrations giving the population the opportunity to accumulate genomic changes, eventually leading to full resistance (Drlica et al. 2008). In clinical practice, bacteria can frequently encounter significantly lower drug concentrations in host-niches such as the nasopharynx, inner ear, or lungs compared to plasma levels (Rybak 2006). Exposure to subinhibitory concentrations of antibiotics may be capable of reducing bacterial growth rates, but can fail to fully eradicate infections, providing selective pressure for acquired resistance. Outside clinical settings, environments containing antibiotics are plentiful, especially due to the rise in antibiotic usage in humans, agriculture, and veterinary medicine (D'Costa et al. 2006; Watkinson et al. 2007). In such environments, selection for antibiotic resistance is likely and frequent (Gullberg et al. 2011).

There are several mechanisms whereby bacteria can resist antibiotic stress including modification of the antibiotic's direct target, enzymatic drug inactivation, and reduction of intracellular drug concentrations via efflux pumps (Walsh 2000; McKeegan et al. 2002; Wright 2003) (Fig. 1). Adaptation—the process by which bacteria attain such mechanisms of resistance—can happen through two modes: horizontal and vertical evolution. The horizontal mode of adaptation (horizontal gene transfer; HGT) involves the acquisition of genetic material from organisms that share the same environment, whereas the vertical mode of adaptation involves the acquisition of *de novo* mutations. Both modes have an important role in shaping the pangenome of bacterial species (Santajit and Indrawattana 2016; Sommer et al. 2017). The use of antibiotics can exert selective pressures that fix horizontally transferred genes or acquired mutations in a population. Examples of HGT include integrons carrying *mecA* which converts methicillin-sensitive *S. aureus* (MSSA) to the resistant "superbug" MRSA (Wielders et al. 2002), beta-lactamases in *P. aeruginosa*, *A. baumannii*, and various species of Enterobacteriaceae (Weldhagen 2004), and macrolide resistance in *Staphylococcus epidermidis* (Lampson et al. 1986) and *Streptococcus pneumoniae* (Chancey et al. 2015). Examples of de novo resistance mutations are plentiful, including mutations in topoisomerase subunits *gyrA* and *parC* conferring resistance to fluoroquinolones (Fàbrega et al. 2009) or in different penicillin-binding proteins, which confer resistance to beta-lactams (Murakami et al. 1987; Sauvage et al. 2002; Munita and Arias 2016; Gifford et al. 2018). Moreover, both modes of evolution can be accelerated by antibiotics. On one hand, fluoroquinolones can induce horizontal gene transfer by activating competence in *S. pneumoniae* (Prudhomme et al. 2006; Slager et al. 2014), while on the other hand, the use of the same class of antibiotics can increase the mutation rate (Lindgren et al. 2003). Importantly, the maintenance of newly acquired resistance in a given population, and its dissemination among species, relies heavily on the associated fitness cost (Melnyk et al. 2015). For instance, the cost of metabolite production in a given reaction may constrain the evolution of antibiotic resistance, highlighting the role of bacterial metabolism and environment on antimicrobial adaptation (Zampieri et al. 2017a). This cost may be different in strains with different

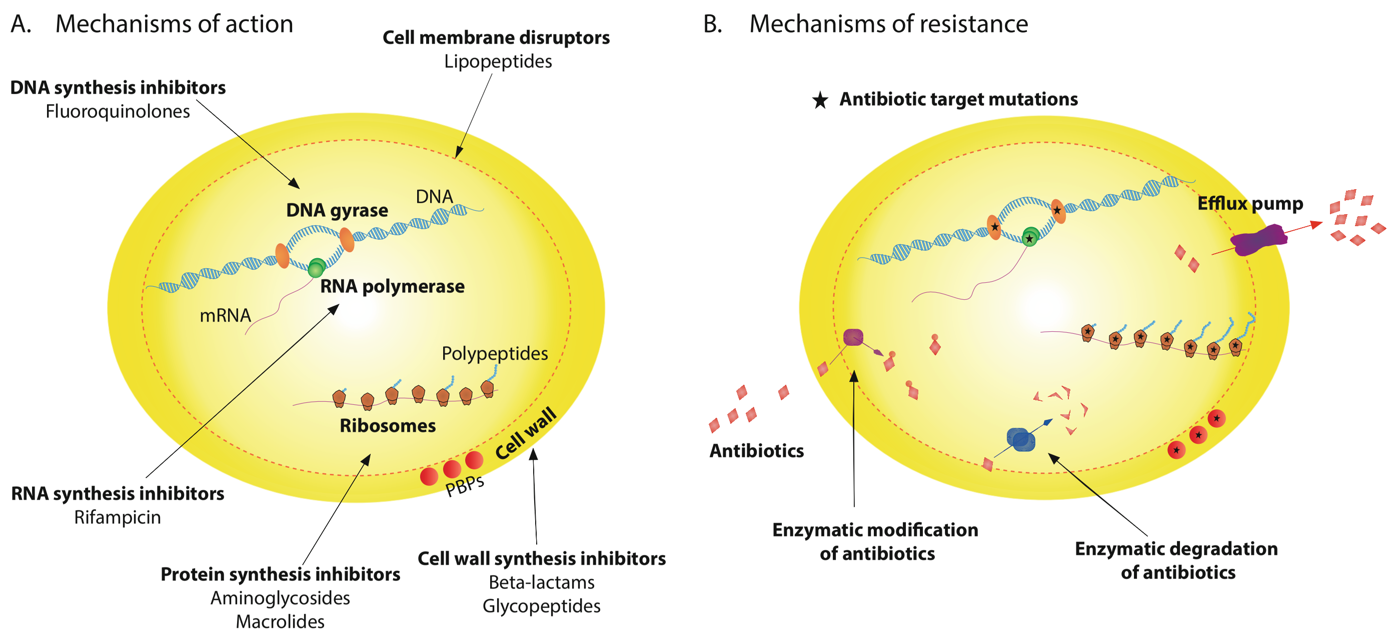


Fig. 1 Action and resistance mechanisms of antibiotics in bacteria. Schematic representation of the most common antibiotic mechanisms of action (**a**) and resistance (**b**) found in bacteria. (**a**) Five major groups of antibiotic classes are represented according to their mechanisms of action: DNA synthesis inhibitors (antibiotics that interfere with DNA replication targeting DNA gyrase and Topoisomerase IV), RNA synthesis inhibitors (antibiotics that interfere with RNA polymerase, thereby inhibiting transcription), Cell wall synthesis inhibitors (antibiotics that interfere in cell wall biosynthesis by inhibition of proteins such as the penicillin-binding protein family), Protein synthesis inhibitors (antibiotics that interact with one of the subunits of the ribosome, thereby interfering with translation), and Cell membrane disruptors (antibiotics that bind or insert into the membrane and cause depolarization). (**b**) The most common mechanisms of resistance among bacteria include acquisition of mutations in the specific antibiotic targets (represented as a black star), active efflux of antibiotics accomplished by integral membrane transporters (efflux pumps), and biosynthesis of enzymes capable of deactivating antibiotics through chemical modification and/or degradation. Yellow ovals: bacterium; brown oval: DNA gyrase; green circles: RNA polymerase; blue double helix: DNA; purple single helix: mRNA, messenger RNA; brown blocks on mRNA: Ribosomes; orange circles: penicillin-binding proteins; light blue circle chains: peptides; red diamonds: antibiotics; yellow border: cell wall; orange dotted border: cell membrane

genetic backgrounds, suggesting that resistance maintenance depends on the bacterial metabolic cost/status as well as for instance the bacterial transcriptional profile in a particular environment (Cornick and Bentley 2012). These negative fitness costs (i.e., reduced growth or replication rates) suggest that in the absence of the antibiotic pressure the adaptive mutations would disappear from the population, nevertheless adapted populations rarely revert to their wild-type versions, and new mutations can compensate for the fitness cost (Andersson and Hughes 2010).

Antibiotics usually target important cellular functions (e.g., cell wall synthesis, DNA replication, or protein synthesis), involving highly conserved and often essential genes present in a wide range of bacteria (Hershberg et al. 2008) (Fig. 1). However, it has become clear that while antibiotics may have very specific targets, the bacterial response to antibiotics and the occurrence of resistance is much more distributed across the genome. For instance, we and others have assayed the antibiotic response of bacteria through genetic perturbations (Fajardo et al. 2008; Tamae et al. 2008; Breidenstein et al. 2008; Schurek et al. 2008; Girgis et al. 2009; Nichols et al. 2011; van Opijnen and Camilli 2012; van Opijnen et al. 2016), which established that a large number of genes and pathways can influence drug susceptibility. These findings underline that we have a limited view of how an antibiotic inhibits a bacterial cell; instead of just a drug–target binary interaction, it is a complex, multifactorial process that begins with that interaction but propagates into various biochemical, metabolic, and regulatory processes of the cell (Tomasz 1979; Vakulenko and Mobashery 2003; Floss and Yu 2005; Drlica et al. 2008; Chandrasekaran et al. 2016; van Opijnen et al. 2016). Thus, a bacterium’s resistance to an antibiotic partially stems from the genome-wide program that is triggered by that antibiotic. This means that small alterations to this program may establish the bacterium on the road to the development of resistance (Albert et al. 2005; El’Garch et al. 2007; Kohanski and Collins 2008; Kohanski et al. 2010; Baquero et al. 2011). So far it has largely been ignored that the genetic diversity present in a pangenome, and the often multiple trajectories that can lead to resistance, can result in strain- and/or species-specific-resistant mechanisms with different fitness costs for maintaining resistance mutations. We believe that all these factors have contributed and are still contributing to the emergence of a diverse resistome, (Davies and Davies 2010; Blair et al. 2015; Munita and Arias 2016) that only makes sense when viewed from a pangenomic context and which makes both the discovery and tracking of resistance as well as the treatment of resistant bacteria far more complex than previously thought.

1.1 Species- and Strain-Specific Differences in Adaptation to Antibiotics

The influence of the pangenome on the complexity underlying the evolution of resistance can be seen both on a between and within species level. For instance, different mechanisms and mutations that trigger resistance to the same antibiotic have mainly been found between species. Additionally, interactions between drug-resistance mutations and genetic backgrounds triggering differences in resistance levels have been found at both the between and within species level. In the following section, we discuss how the pangenome, or genetic differences between strains and species, affect the mechanism and/or the level of resistance that evolves.

1.1.1 Species-Specific Resistance: There Is More Than One Way to Become Resistant

Due to specific (pangenomic) genetic characteristics, different species can adopt different mechanisms of adaptation to become resistant to the same antibiotic. Additionally, different species can acquire different mutations or genes to achieve the same resistance mechanism to the same antibiotic. A well-documented example of the first scenario is beta-lactam-resistant mechanisms among Gram-negative and Gram-positive bacteria. Beta-lactam antibiotics inhibit bacterial cell wall synthesis by targeting penicillin-binding proteins (PBPs), a group of enzymes that are present in all bacterial species and which catalyze peptidoglycan cross-linking. PBPs interact with beta-lactams via an active site serine and form a relatively stable covalent complex (Sibold et al. 1994). The primary resistance mechanism against clinically important beta-lactams (e.g., penicillin, carbapenem, cephalosporin) is different between Gram-negative and Gram-positive bacteria. In Gram-negative bacteria, beta-lactam resistance is commonly driven by the acquisition of hydrolyzing beta-lactamases that inactivate the drug. In contrast, beta-lactam resistance in most Gram-positive species is mediated by target modifications, with the exception of staphylococcal penicillinase (Rosdahl 1985; Skov et al. 1995). For instance, beta-lactam-resistant *Enterococcus faecium* have acquired mutations in an essential PBP (PBP5) that reduce the accessibility of the active site and result in a low-affinity form (PBP5fm) (Sauvage et al. 2002). Similar low-affinity PBPs have also been reported in methicillin-resistant *S. aureus* (MRSA) (Murakami et al. 1987) and in beta-lactam-resistant strains of *S. pneumoniae* (encoded by mosaic genes acquired through HGT) (Sibold et al. 1994; Reichmann et al. 1996). It seems likely that this divergence in beta-lactam-resistant mechanisms between Gram-negative and Gram-positive bacteria arose from the differences in their cell envelopes (Munita and Arias 2016). In Gram-negative bacteria, the presence of an outer membrane and associated porins allows for the entry and accumulation of beta-lactams in the periplasmic space, prior to binding PBPs in the inner membrane (Fig. 2a). Such compartmentalization allows for beta-lactamase accumulation at sufficient concentrations and effective deconstruction of the beta-lactam molecules.

While species-specific differences in antibiotic resistance can come from very different mechanisms, there are examples where the target is the same, but the manner in which it is targeted is different. For example, macrolides target the peptidyl site of nascent peptides in the large subunit of bacterial ribosomes, thereby inhibiting protein synthesis. Many cases of clinical macrolide resistance are caused

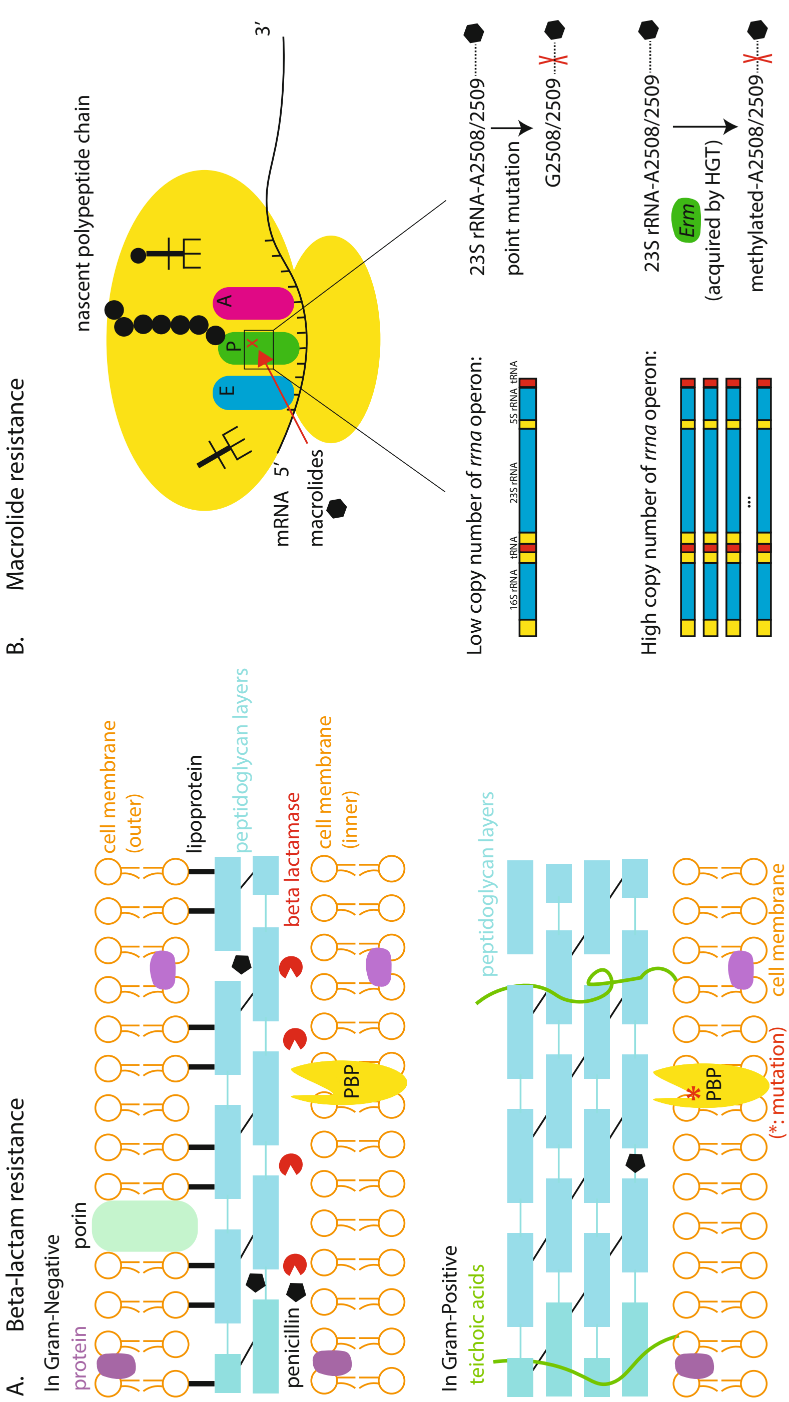

A. Beta-lactam resistance
In Gram-Negative
protein
porin
cell membrane (outer)
lipoprotein
peptidoglycan layers
beta lactamase
penicillin
PBP
cell membrane (inner)
In Gram-Positive
teichoic acids
peptidoglycan layers
PBP
(*: mutation)
cell membrane
B. Macrolide resistance
nascent polypeptide chain
E
P
A
mRNA 5'
3'
macrolides
Low copy number of *rrna* operon:
16S rRNA
tRNA
23S rRNA
5S rRNA
tRNA
High copy number of *rrna* operon:
...
23S rRNA-A2508/2509
point mutation
G2508/2509
23S rRNA-A2508/2509
Erm
(acquired by HGT)
methylated-A2508/2509

Fig. 2 Species-specific antibiotic-resistance mechanisms. (**a**) In Gram-negative bacteria, acquisition of beta-lactamase is the preferred mechanism of beta-lactam resistance. This is most likely due to the presence of a periplasmic space that allows for beta-lactamase accumulation, resulting in effective degradation of the antibiotic. In Gram-positive bacteria, beta-lactam resistance is commonly driven by mutations in antibiotic target, penicillin-binding proteins (PBPs), which reduces the beta-lactam-binding affinity. (**b**) Macrolides target the peptidyl (P) site of ribosomes and interact with nucleotides A2508/2509 in the 23S rRNA. Macrolide resistance is commonly driven by modifications of target nucleotides. In species with low copy number of the *rrna* operon, macrolide resistance via 23S rRNA modification is frequently achieved by point mutations of A2508/2509. In species with high copy number of the *rrna* operon, target nucleotide modification is commonly achieved by *erm*-methylation

by mutations at specific nucleotide positions in the 23S rRNA. Due to differences in the copy number of the ribosomal RNA operon (*rrna*), different species have been shown to have different macrolide-resistant mutations in the 23S rRNA gene (Fig. 2b). Generally, mutations at A2058 or A2059 in the 23S rRNA (using *E. coli* nucleotide sequence numbering) confers macrolide resistance for many pathogenic bacteria, predominantly bacteria with one or two copies of the *rrna* operon, such as azithromycin-resistant *Treponema pallidum* (Stamm and Bergen 2000; Matejkova et al. 2009), clarithromycin-resistant *Mycobacterium* species (Meier et al. 1994; Nash and Inderlied 1995; Wallace et al. 1996), and *Helicobacter pylori* with resistance to macrolide–lincosamide–streptogramin B antibiotics (termed as the MLS_B phenotype) (Wang and Taylor 1998). A2058 and/or A2059 mutations change the structure of the drug-binding pocket and thereby reduce the binding affinity of the drug contributing to resistance. In bacteria with higher copy numbers of *rrna*, such as *Staphylococcus, Enterococcus*, and Streptococcus, acquisition of point mutations on all or multiple copies of the 23S rRNA genes is highly improbable. Instead, macrolide resistance via 23S rRNA modification is frequently achieved by *erm*-methylation of target nucleotides. *Erm* genes are mobile genes that encode 23S rRNA methylases and can catalyze dimethylation of A2058 (Toh et al. 2007). In *S. pneumoniae*, ErmB provides a high level of resistance to erythromycin (MIC $>$ 256 μg/mL) (Schroeder and Stephens 2016), which suggests that resistance level conferred by the same mutation is also dependent on the genetic background.

1.1.2 Interactions Between Resistance Mutations and Genetic Background Can Affect the Level of Resistance

While it may not come as a complete surprise that different species can adopt different strategies to overcome resistance, recent studies have shown that when different species or strains do have the same strategy to become resistant, the same mutation does not automatically result in the same level of resistance. This can be caused by differences in the genetic background and is a good example of how genetic differences between species and strains, can have important effects on (the emergence of) resistance. Examples at the species level are loss-of-function mutations of the 16S rRNA-specific methyltransferase GidB involved in streptomycin resistance (Okamoto et al. 2007; Koskiniemi et al. 2011). Streptomycin, an aminoglycoside antibiotic, binds to the 30S subunit of the ribosome and causes misreading of the correct tRNA. These mutations have been identified in low-to-intermediate levels of streptomycin resistance in multiple bacteria, such as *M. tuberculosis*, *Mycobacterium smegmatis*, *S. aureus*, and *E. coli* (Okamoto et al. 2007; Wong et al. 2011; Perdigao et al. 2014). Koskiniemi et al. (2011) showed that high-level streptomycin resistance caused by the loss of GidB is largely dependent on the presence of an aminoglycoside adenyltransferase (AadA) in the bacterium's genome, which is an enzyme that modifies and thereby inactivates aminoglycosides (Tait et al. 1985; Svab et al. 1990; Magrini et al. 1998; Frank et al. 2003). In an experimental evolution study, streptomycin-adapted *Salmonella typhimurium* strains that have both the *aadA* gene and *gidB* mutations gained a higher level of streptomycin resistance than strains having either one alone (Wistrand-Yuen et al. 2018). Species with *aadA*, e.g., *S. typhimurium*, can thereby gain a high level of streptomycin resistance, while species that lack this enzyme (e.g., *E. coli*, *S. aureus*, and *M. tuberculosis*) only obtain low-level resistance (Okamoto et al. 2007).

Within a species, the same mutations may also not necessarily result in the same level of resistance. One such example is resistance in *M. tuberculosis* to isoniazid (INH). As a prodrug, INH must be processed by the mycobacterial enzyme KatG into its active form, isonicotinic acyl-NADH. The active drug then binds the enoyl-acyl carrier protein reductase InhA and blocks the synthesis of mycolic acid (Quémard et al. 1991). In *M. tuberculosis*, the primary INH-resistance mechanism is via a point mutation in KatG (e.g., S315T), which results in a partially active protein that reduces INH binding while retaining enough activity to support bacterial survival. Another frequently observed resistance mutation is in the promoter region of the target gene *inhA* (Lee et al. 2001). Strains that have *inh* promoter mutations have been observed to show different levels of INH resistance based on their phylogenetic lineages. *M. tuberculosis* is grouped in six main phylogenetic lineages (Hershberg et al. 2008; Comas et al. 2010): three modern lineages that have evolved in regions with high-density populations and recent massive demographic expansion (i.e., lineage 4: Europe and America, lineage 3: India and East Africa, lineage 2: East Asia) and three ancient lineages from older and low-density populations (i.e., lineage 1: the Philippines, lineage 5: Rim of Indian ocean, and lineage 6: west Africa) (Portevin et al. 2011). A study of 158 isolates of multidrug-resistant *M. tuberculosis* revealed that mutations in the *inhA* promoter cause high level of INH resistance ($\geq$3.0 μg/mL) only in the modern lineages 2 and 3, while these mutations cause low-level resistance (MIC $<$3.0 μg/mL) mainly in ancient lineages 1 and 5 (Fenner et al. 2012). Although *M. tuberculosis* harbors limited genetic diversity compared to other species, multiple studies have suggested that the variation in drug-resistant phenotypes of *M. tuberculosis* could be at least partially explained by epistatic interactions among the genetic background of different phylogenetic lineages, compensatory mutations and drug-resistance mutations (Gagneux et al. 2006; Fenner et al. 2012; Gygli et al. 2017).

1.1.3 The Ability of Evolving Antibiotic Resistance May Vary Across Species Due to Epistatic Interactions and/or "Potentiator" Genes

Apart from epistatic interactions between genetic background and drug-resistance mutations, the presence of potentiator genes can make it possible for a novel trait to evolve that would otherwise be inaccessible (Blount et al. 2012; Lind et al. 2015). Depending on the genetic background, the presence of potentiators of antibiotic-resistance genes can prime strains to evolve resistance. To uncover the role of potentiators in different genetic backgrounds, Gifford and colleagues evolved eight strains in the *Pseudomonas* genus to the beta-lactam antibiotic ceftazidime and compared their pathways that led to resistance (Gifford et al. 2018). Their results show that *Pseudomonas* species that have the transcription factor *ampR* (*P. protegens* and *P. fluorescens*) evolve ceftazidime resistance faster than species lacking this gene (*P. mendocina* or *P. fulva*) (Gifford et al. 2018). AmpR has been shown to increase the expression of beta-lactamase *ampC* upon the inactivation of peptidoglycan synthesis (Mark et al. 2011; Ropy et al. 2015). The authors hypothesized that *ampR* potentiates ceftazidime adaptation by allowing mutations in peptidoglycan biosynthesis genes such as *ampD*, *pml*, and *dacB*. Indeed, in *P. aeruginosa*, *dacB* inactivating mutations have only been observed in genetic backgrounds harboring *ampR* (Moya et al. 2009; Mark et al. 2011). These findings show that (clinically relevant) high-resistance level markers (e.g., mutations, genes acquired by HGT) should be considered and validated in different genetic backgrounds, and thus in a pangenomic context.

1.2 Strain- and Species-Specific Phenotypic Stress Responses to Antibiotics

Recent advances in high-throughput techniques involving mutant libraries as well as various omics approaches have allowed for unprecedented understanding of how bacteria respond to antibiotic-mediated stress. Such strategies have shown the diversity of antibiotic responses within species represented by a large pangenome as well as between species. Various examples discussed below show that antibiotics can induce stress throughout the bacterium both at the direct target of the antibiotic as well as at off-target pathways throughout the genome. Due to the pangenome and the consequent differences in genetic backgrounds, strains and species respond to antibiotics with (slightly) different sets of genes and thereby experience antibiotic stress in different ways. This means the selective pressures a bacterium experiences can be strain and/or species specific and drive the evolution of resistance in a strain- or species-specific manner. As a result, the pangenome not only affects the manner in which stress is experienced, but that same stress (e.g., antibiotics) also contributes to maintaining and expanding the pangenome.

1.2.1 High-Throughput Tools for Investigating the Bacterial Response to Stress

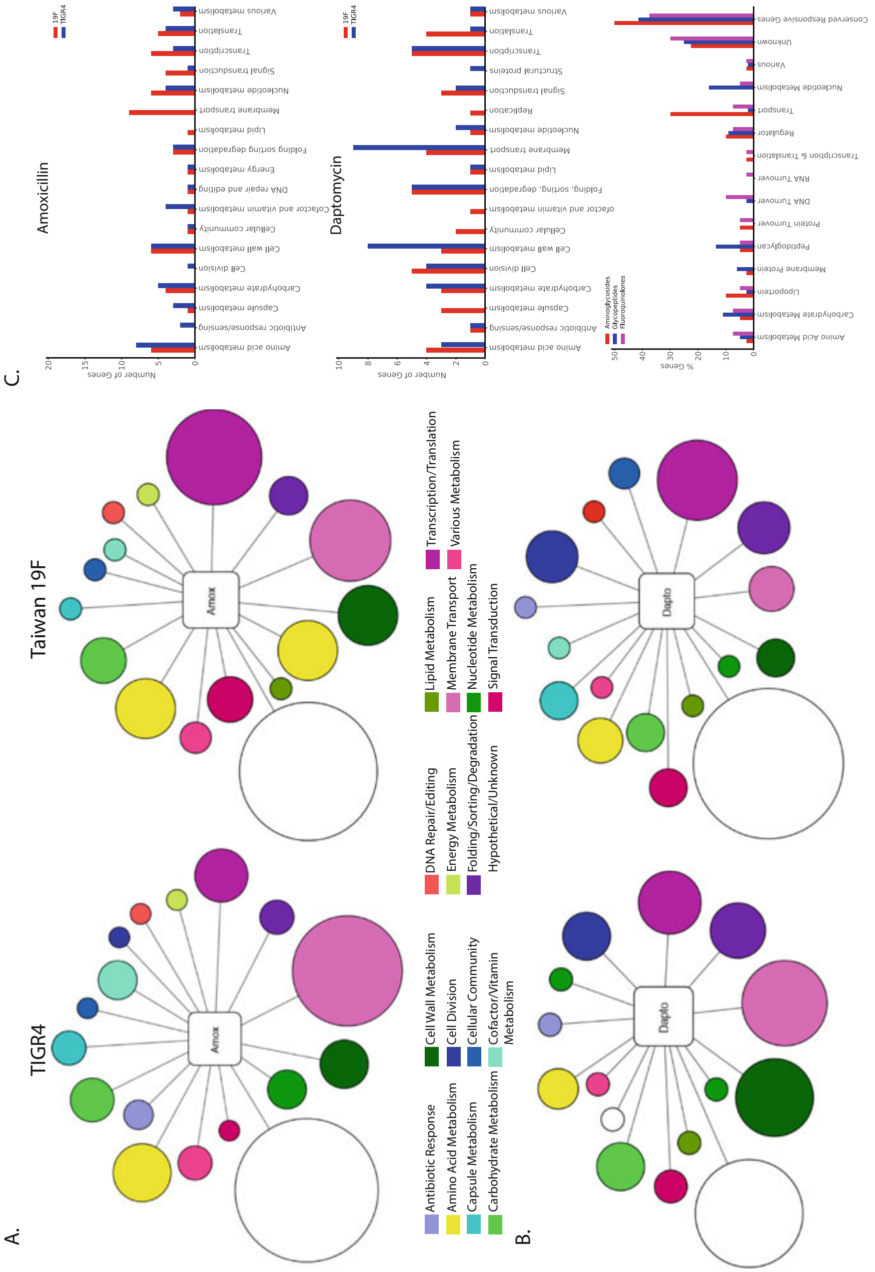
With the rise of low-cost sequencing options, whole genome sequencing (WGS) has proved useful for identifying antibiotic-resistant bacteria by looking for the presence of certain genes (e.g., efflux pumps), insertion–deletions, and other polymorphisms associated with antibiotic resistance (Boissy et al. 2011; Zankari et al. 2012; Liu et al. 2014; McDermott et al. 2016; Zeng et al. 2018). The increased availability of large collections of bacterial whole genome sequences has allowed the identification of numerous single nucleotide polymorphisms (SNPs) associated with drug resistance through genome-wide association studies (Power et al. 2017). Resistance-associated SNPs have been identified for a number of pathogenic bacteria including *M. tuberculosis* (Desjardins et al. 2016)*, S. pneumoniae* (Chewapreecha et al. 2014), and *S. aureus* (Alam et al. 2014). For antibiotic surveillance, the ability to identify features such as SNPs means WGS provides much more detailed information compared to traditional phenotyping such as multilocus sequence typing (MLST). This increased resolution can be used to predict antibiotic resistance for clinical isolates based on databases of known antibiotic-resistance determinants (Sandgren et al. 2009; McArthur et al. 2013; Stoesser et al. 2013; Walker et al. 2015; Lakin et al. 2017). Recent work has even demonstrated the ability to identify resistant strains as the sample is sequenced (Břinda et al. 2018) potentially leading to point-of-care devices which can guide appropriate use of antibiotics by clinicians. Nevertheless, predictions of resistance are limited to the antibiotics that have been previously tested (such as clinically important first- and second-line antibiotics), which hampers their utility in predicting bacterial responses to novel antibiotics. Thereby, WGS provides a snapshot of the presence or absence of resistance determinants but cannot directly provide information on what genes or pathways are involved in responding to the stress induced by antibiotics. Consequently, while WGS and MLST are highly useful for resistance surveillance and may guide treatment options, they are more limited in their ability to tease apart phenotypic responses to antibiotics for the purpose of understanding and potentially predicting how resistance develops.

In contrast, the use of ordered mutant libraries can directly link genes to observed phenotypes (Jacobs et al. 2003; Baba et al. 2006), which have allowed the detailed characterization of how bacteria respond to various antibiotics (Nichols et al. 2011). However, these libraries are limited by being time consuming to construct, making it less amenable for a wide variety of bacteria. The advent of techniques such as Tn-Seq (van Opijnen et al. 2009), INSeq (Goodman et al. 2009), HiTS (Gawronski et al. 2009) and TRADiS (Langridge et al. 2009), and variants like RB-TnSeq (Wetmore et al. 2015; Price et al. 2018) and droplet Tn-Seq (Thibault et al. 2019) offer a high-throughput alternative which is easily adaptable. In general, all these techniques rely on generating transposon-insertion libraries, which can be assayed by high-throughput sequencing for the relative frequency of mutants grown in a particular stress-inducing environment such as subinhibitory concentrations of antibiotics. In this way, the phenotype of each genetic mutant can be determined, showing directly how bacteria respond to antibiotics and the genes that benefit or hinder the bacteria's ability to respond to this stress. Thanks to a diverse number of transposon systems and the relative ease of creating mutant libraries, these techniques are amenable to a wide variety of bacterial species and individual strains, providing data within the context of the genetic background of each assayed strain. Characterization of the response to antibiotics can also be complemented by various "omic" approaches. These include transcriptomic (Jensen et al. 2017; Qin et al. 2018), metabolomic (Zampieri et al. 2017b), and proteomic (Pérez-Llarena and Bou 2016; Ma et al. 2017) analyses. The datasets generated by these techniques can also be overlaid with one another to provide a holistic understanding of how bacteria respond to antibiotic stress (Jensen et al. 2017).

Studies utilizing Tn-Seq and related methods have shown that antibiotic-induced stress involves the target of the antibiotic and also extends throughout the entire genome of the bacterium. For example, fluoroquinolones like ciprofloxacin, levofloxacin, and norfloxacin target topoisomerase IV and DNA gyrase, critical enzymes utilized in DNA synthesis. In the Gram-positive *S. pneumoniae* and the Gram-negative *A. baumannii*, Tn-Seq profiles for fluoroquinolones show that genes involved in DNA replication and repair such as *recN* and *xseA* are important for responding to these antibiotics. While these genes are not direct targets, the inhibition of DNA replication by targeting gyrase and topoisomerase triggers DNA damage and thus explains the indirect importance of genes involved in DNA repair (van Opijnen and Camilli 2012; Geisinger et al. 2019). In addition, Tn-Seq profiles show a role for genes even beyond those related to DNA repair and replication and indicate the importance of genes with diverse functions including amino acid and carbohydrate metabolism. In *P. aeruginosa*, the aminoglycoside tobramycin also involves a diverse number of responsive genes, including those involved in cell division, carbohydrate metabolism, and membrane metabolism (Gallagher et al. 2011). Similar findings can be observed in data from *E. coli* where colony sizes were measured for an ordered mutant library grown in the presence of various stressors, including antibiotics (Nichols et al. 2011). For example, a screen with trimethoprim/sulfamethoxazole, which targets the folate biosynthesis pathway shows an important role for genes involved in this pathway, including *mogA* and *folM,* as well as genes involved in nucleotide metabolism. But again, responsive genes also include those involved in carbohydrate metabolism, glycan biosynthesis, and membrane transport. These examples highlight that while stress may be felt acutely at the antibiotic's target, it extends beyond the primary target and results in selective pressures acting throughout the genome. The importance of this is further confirmed by the observation that resistant clinical isolates often have mutations at sites throughout the genome that resolve such stress and/or work in a compensatory manner (Albert et al. 2005; El'Garch et al. 2007). Interestingly, targeting genes involved in off-target responses can create an opportunity for therapeutic intervention by generating synergy between the off-target gene/response and the assayed drug.

1.2.2 Strain-Specific Responses to Antibiotic Stress

In addition to showing that stress can reverberate throughout the genome, Tn-Seq is able to reveal how the genetic background of a strain affects the response to antibiotic stress. Several examples have shown that the genes and pathways involved in responding to antibiotic stress can be strain specific. For instance, *S. pneumoniae* strains TIGR4 and Taiwan-19F are similarly susceptible to daptomycin, however, Tn-Seq results show that only 50% of the genes responding to daptomycin are common to both strains, with the other 50% being strain specific (van Opijnen et al. 2016) (Fig. 3). Moreover, the distribution of the functional categories of the responsive genes is significantly different between the two strains. This lack of


A.
TIGR4
Taiwan 19F
Amox
Amox
Antibiotic Response
Amino Acid Metabolism
Capsule Metabolism
Carbohydrate Metabolism
Cell Wall Metabolism
Cell Division
Cellular Community
Cofactor/Vitamin Metabolism
DNA Repair/Editing
Energy Metabolism
Folding/Sorting/Degradation
Hypothetical/Unknown
Lipid Metabolism
Membrane Transport
Nucleotide Metabolism
Signal Transduction
Transcription/Translation
Various Metabolism
B.
Dapto
Dapto
C.
Amoxicillin
19F
TIGR4
Number of Genes
0
5
10
15
20
Amino acid metabolism
Antibiotic response/sensing
Capsule metabolism
Carbohydrate metabolism
Cell division
Cell wall metabolism
Cellular community
Cofactor and vitamin metabolism
DNA repair and editing
Energy metabolism
Folding sorting degradation
Lipid metabolism
Membrane transport
Nucleotide metabolism
Signal transduction
Transcription
Translation
Various metabolism
Daptomycin
19F
TIGR4
Number of Genes
0
2
4
6
8
10
Amino acid metabolism
Antibiotic response/sensing
Capsule metabolism
Carbohydrate metabolism
Cell division
Cell wall metabolism
Cellular community
ofactor and vitamin metabolism
Folding, sorting, degradation
Lipid metabolism
Membrane transport
Nucleotide metabolism
Replication
Signal transduction
Structural proteins
Transcription
Translation
Various metabolism
Aminoglycosides
Glycopeptides
Fluoroquinolones
% Genes
0
10
20
30
40
50
Amino Acid Metabolism
Carbohydrate Metabolism
Lipoporein
Membrane Protein
Peptidoglycan
Protein Turnover
DNA Turnover
RNA Turnover
Transcription & Translation
Regulator
Transport
Nucleotide Metabolism
Various
Unknown
Conserved Responsive Genes

conserved response is also observed for antibiotics representing fluoroquinolones, aminoglycosides, and glycopeptides, with only 40–50% of the responsive genes conserved between these two strains for a particular antibiotic. Nevertheless, when the functional categories are combined into larger groupings corresponding to different domains of the cell's physiology, there is no difference in the distribution between the two strains. This suggests that despite strain-specific differences in response at the gene level, the global response is more similar (van Opijnen et al. 2016).

In *Mycobacterium tuberculosis,* in vitro Tn-Seq experiments have shown that several clinical strains have an increased requirement for the gene encoding KatG, compared to reference strain H37Rv (Carey et al. 2018). As discussed, KatG is an activator of the first-line *M. tuberculosis* antibiotic isoniazid, and adaptation experiments have shown that loss-of-function mutations in *katG* can result in isoniazid resistance. However, such mutations occur at a low frequency in clinical strains (Gagneux et al. 2006; Vilchèze and Jacobs 2014), which suggests that the increased fitness cost of mutating *katG* in clinical strains decreases the frequency of acquisition of isoniazid mutants compared to H37Rv. Furthermore, Tn-Seq identified minimal fitness costs for losing *glcB* (a maleate synthase involved in the glyoxylate shunt, which is important for carbon and fatty acid metabolism) in some clinical strains, whereas it is highly important in other strains (Carey et al. 2018). The authors hypothesized that such differential requirements for *glcB* would result in correspondingly differential responses to a novel inhibitor of this protein. Indeed, they found that strains showing less of a requirement for *glcB* are less susceptible to the inhibitor. This type of variability illustrates how the pangenome affects responses and consequently adaptive solutions to antibiotic stress and underscores why therapies may not produce consistent results across all strains. Furthermore, the finding that strains can demonstrate considerable variation in their response to antibiotics underscores how caution must be taken when evaluating studies that are based on a single strain and thereby ignore differential responses that may be present throughout the pangenome.

Fig. 3 Strain-specific differences in responses to the same antibiotic. Networks show the relative number of responsive genes of a given functional category responding to either amoxicillin (**a**) or daptomycin (**b**) for *S. pneumoniae* strains TIGR4 and Taiwan 19F. The number of genes for each group is shown in the charts on the right side. Note the diversity of functional categories beyond the membrane target of both antibiotics. Each strain also responds to the antibiotics with slightly different functional categories. While the strains appear to lack genetic and functional conservancy, they do respond globally in the same way, when the functional categories are condensed into categories involving the capsule, membrane, cellular control, and metabolism. (**c**) The functional categories show a similar diversity of functions when responding to aminoglycosides, glycopeptides, and fluoroquinolones for both TIGR4 and 19F

1.2.3 Gene Homology Frameworks to Uncover Differential Responses Across Bacterial Species

In addition to considering strain-specific responses to the same antibiotic, we have assessed stress responses at the species level to determine how similar antibiotic response patterns are. By combining data from a variety of sources (Nichols et al. 2011; Murray et al. 2015) and generating two frameworks utilizing the OMA and PATRIC databases (Wattam et al. 2017; Altenhoff et al. 2018) the responses of *E. coli*, *P. aeruginosa*, *A. baumannii*, and *S. pneumoniae* to ciprofloxacin, could be compared. While not all responsive genes have homologs in all species, a consistent pattern is observed for these diverse species. Genes involved in DNA replication and repair such as *recN* and *xseA* are important in all four species and in both homology frameworks. Additional nonhomologous genes annotated as involved in DNA replication and repair are also observed in each of the four species. Each species also has responsive genes that are involved in various metabolic functions and cellular processes not related to DNA repair. Nevertheless, pairwise strain comparisons indicate only 5–10% of homologs not involving DNA replication and repair are shared between one or more species. This suggests that species may respond to antibiotic stress similarly at the antibiotic target and related pathways, but individually each species is responding with a unique program depending on its genetic background and thus coping with unique selective pressures that can influence the emergence of resistance.

1.3 The Role of the PanGenome in Predicting Adaptation to Antibiotic Stress

We have so far discussed that mutations responsible for the resistance phenotype to a certain antibiotic can either be common or be specific to a strain or a species. Nevertheless, the types of adaptive-resistance mutations, and the order in which they arise and fix in a population (adaptive trajectories) have shown to be replicable (Elena and Lenski 2003), which suggests adaptive evolution is, at least to a certain extent, constrained. In this section, we argue that the genetic background and the environmental context are two major factors that constrain adaptive evolution (e.g., during adaptation to antibiotics). We first discuss the role of genetic interactions and how they reduce the number of available adaptive trajectories, and propose a pangenome-wide view of studying genetic interactions. Next, we discuss the possibility of using how a selective pressure in the environment is sensed and experienced by an organism (environmental context) to predict where on the genome adaptive mutations will appear when the selective pressure is maintained. We argue that the predictions based on environmental context can be improved by the addition of pangenome-related information, such as the conservation of genetic sequences across many related organisms.

1.3.1 Adaptive Evolution Is Replicable, Therefore Predictable

The analysis of sequence sets on a pangenome scale allows associations to be made between genetic changes and antibiotic-resistant phenotypes. Such pangenome-wide association studies have revealed common sets of mutations that appear in organisms resistant to a certain antibiotic (Croucher et al. 2011; Mobegi et al. 2017a; Del Barrio-Tofiño et al. 2017). Moreover, phylogenetic reconstructions suggest the same resistance-causing mutations have appeared independently, and multiple times in geographically separated strains (Croucher et al. 2011; Farhat et al. 2013; Chewapreecha et al. 2014). While, such ad hoc associations have the power of explaining the genetic basis of a certain phenotype, they rarely offer a predictive model for future adaptive trajectories. However, the observation that the same mutations have appeared in different pathogens independently suggests that adaptive evolution is not an entirely random process. This can be further seen in lab-directed adaptation experiments, where common sets of mutations keep reappearing in independent populations under the same selective pressure (Lang et al. 2013). These common adaptive trajectories demonstrate the replicability of adaptive evolution, which is not to say evolution is an entirely deterministic process. The emergence of new sequence variants is stochastic and phenomena such as hitchhiking genetic regions, genetic drift, and clonal interference can incorporate different degrees of randomness influencing which mutations will reach fixation and how (Lang et al. 2013). Yet, the replicability of adaptive evolution in antibiotic resistance suggests that there are a limited number of adaptive trajectories available to the adapting organism. In other words, while there are many possible ways a set of resistance mutations can reach fixation, the majority of those trajectories are not plausible because they are constrained by the environment and the genetic context. This means that if the environmental and genetic constraints a bacterium evolves under can be understood and/or (experimentally) captured, adaptive evolution should become predictable.

1.3.2 Genetic Constraints on Adaptation

In order to understand the genetic constraints on adaptation we need to consider epistatic interactions within the genome. Epistatic interactions are defined as the nonadditive effects of combinations of mutations. For instance, mutations can have different effects on fitness, depending on the genetic background of the organism, i.e., what other mutations are already present in the parental strain (Vogwill et al. 2016). This is well illustrated by experiments that compared the fitness of single mutants with combinations of those singlets into double and triple mutants, where the fitness of the double and triple mutants differed from what is expected under the multiplicative model (i.e., in the absence of epistasis, the fitness of combining mutant A and mutant B in the same genome = fitness of A x fitness of B) (Weinreich et al. 2006; Angst and Hall 2013; Hall and MacLean 2016). Moreover, epistasis influences the order in which mutations appear. If two mutations do not interact with each other, then any order in which they appear is equally likely. However, when mutations do interact, the appearance of one may limit the appearance of the other. For instance, when the interactions of 5 mutations that confer resistance to cefotaxime were mapped out in *E. coli*, only 10 trajectories (out of 120 possible ones) turn out to have non-negligible probabilities of being observed (Weinreich et al. 2006). Another example of this is where lab adaptation of a beta-lactamase in *E. coli* is limited in its trajectory and will follow a certain path (that with the highest likelihood) when a specific initial mutation is present (Salverda et al. 2011).

Since epistatic interactions limit the adaptive trajectories to a few likely ones, mapping out epistatic interactions can help determine which trajectories are most plausible, and thereby contribute to predictions of adaptive evolution. A high-throughput way of determining epistatic interactions is using genome-wide double-knockout screens, as has been done extensively in *Saccharomyces cerevisiae* (Tong et al. 2004), and *Schizosaccharomyces pombe* (Roguev et al. 2008). In these studies, synthetic lethality, which is an "extreme" form of epistasis, was used to build genome-wide epistasis or genetic interaction networks. These networks show the prevalence of epistatic interactions throughout the entire genome, with most genes interacting with at least one other gene, and a few hubs with numerous interactions. A comparison of genetic interaction networks from *S. cerevisiae* (Tong et al. 2004), and *S. pombe* (Roguev et al. 2008) demonstrate that the same interactions are not always present, even when considering genes common to both organisms. In other words, while some interactions are conserved, others may be present or absent, depending on the genetic background. This means that the genetic interaction network of a single organism is not representative of a pangenome-wide genetic interaction network. Therefore, predictions of adaptation based on a single-strain network will be limited to that organism.

1.3.3 A Pangenome-Wide View of Epistasis May Enhance Predictions

Epistatic interactions are more than a collection of gene- or locus-pairs, but rather form a complex network that has both components that are universally true (those interactions that are strain or species independent), and components that are only present in a certain strain or species. When epistatic interactions (on a gene level) are mapped for a single strain, the interactions are limited to the genes present in this one strain. However, the lack of a gene is not equivalent to the lack of the *influence* of that gene. The fact that the gene is absent may actually affect the fitness of strains with this particular genetic background. It is possible that genetic elements that vary considerably in their presence or absence across different strains interact epistatically. Such interactions have indeed been demonstrated between chromosomal mutations and plasmids (Silva et al. 2011) or mobile elements (Stoebel and Dorman 2010), or even between two plasmids (San Millan et al. 2014). Therefore, studies mapping out genetic interactions in a single strain or species can be limiting, showing only the components of a network that applies to the organism being studied. In order to get a comprehensive view, it is necessary to construct a pangenome-wide genetic interaction network.

One possibility is to map out genetic interaction networks on hundreds of related strains/species. However, even with high-throughput screening methods, this approach is limited by time and cost. A more feasible alternative would be inferring genetic interactions through in silico analysis of a collection of genomic sequences. In a simple model, one can assume there is an underlying network of epistatic interactions, where each gene's state (present or absent) influences the states of the genes it is interacting with [analogous to the Ising model describing particle spin states in statistical mechanics (Ising 1925)]. Each viable organism can then be described as a configuration—the presence and absence state of all genes in the pangenome. In this model, the underlying genetic interaction network results in some configurations being more likely than others. It is reasonable to assume that viable organisms are the more likely configurations. Based on this assumption, and considering the observed states of each gene from many genomes, it is possible to infer the underlying network connectivity between genes (Bresler 2014) and identify interactions between genes that are more likely to be universally true, and not strain/ species specific. Such a comprehensive genetic interaction network should give a much better idea about pangenome-wide constraints on adaptive evolution. In fact, the fitness landscape (a popular visual metaphor for the effect of genotypes on fitness) is a pangenomic concept. This long-standing visual lays-out the possible genotypes of an organism (or the existing genotypes in a pangenome) on a flat horizontal surface, and the fitness of each genetic variant is plotted on the vertical axis. Thus, because the fitness landscape considers many genomic variants at once, it inherently represents a pangenome view of fitness. The classical view of the fitness landscape is that there is a single peak of fitness, and an organism adapting under a selective pressure climbs this fitness peak as it accumulates mutations. However, increasing numbers of epistatic interactions result in the fitness landscape becoming decorated with peaks and valleys, forming a rugged surface (Kauffman and Weinberger 1989). This apparent increase in complexity may also explain certain strain-specific adaptive outcomes, as it becomes clearer where local fitness maxima and minima are situated on the landscape. Consequently, the consideration of the pangenome (rather than single genomes) should uncover a comprehensive genetic interaction network.

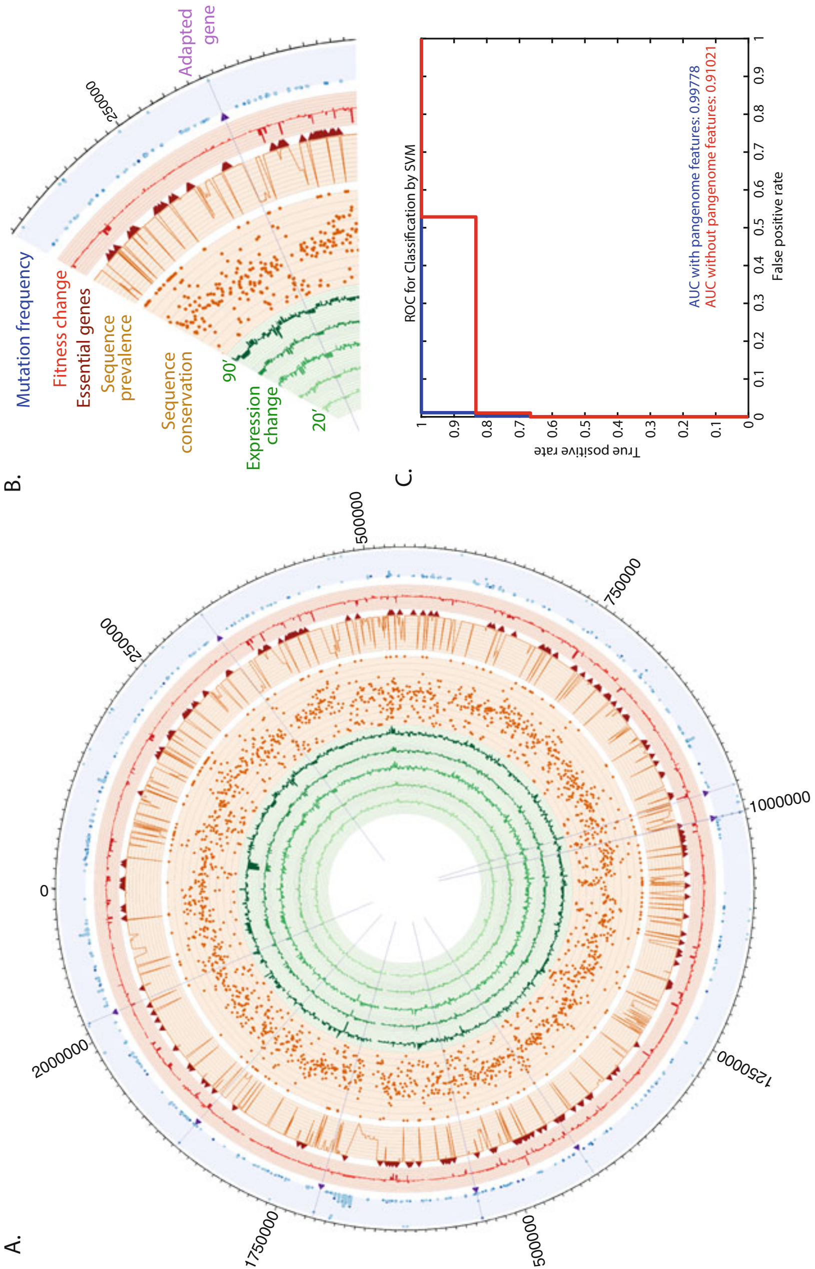
1.3.4 Toward Predicting Adaptive Evolution and the Importance of Pangenomic Information

The fitness landscape has long been considered a constant and rigid surface for each organism. However, genotype is not the only determinant of fitness—the same organism's fitness varies in different environments. Thus, the fitness landscape is a much more fluid concept, and its shape/contour depends on environmentally determined selective pressures. In other words, in addition to the genetic context determining fitness outcome and constraining adaptive evolution, environmental context also plays an important role.

Similar to how genetic interaction networks can reveal genetic constraints on adaptive evolution, multi-omic profiling reveals environmental constraints on adaptive evolution. The manner in which a bacterium will adapt to the environment it finds itself in is linked to how a stress, i.e., a selective pressure (e.g., antibiotic), is sensed and processed by the bacterium (Zhu et al. 2018, 2019). The use of multi-omic profiling (e.g., via Tn-Seq, RNA-Seq) can reveal which genomic loci respond to and are important in overcoming stress in the environment. For instance, Tn-Seq experiments identify the genes that contribute to fitness (phenotypically important genes, or PIGs) under the stress, and RNA-Seq experiments reveal transcriptionally important genes (or TIGs) responding to the stress. A simple assumption would be that because PIGs and TIGs are relevant in the organism's response to stress, they will also be implicated in resolving this stress over the course of adaptive evolution. In other words, genes that acquire adaptive mutations will be TIGs and/or PIGs. While in some cases, PIGs and/or TIGs acquire adaptive mutations, not all adapted genes are PIGs or TIGs (Fig. 4). This, along with the transcriptional and phenotypic responses often involving many different cellular functions, makes it challenging to find straightforward rules that predict which genes will adapt. This has motivated the use of machine learning algorithms to detect potentially multifactorial and complicated determinants of adaptive evolution (Zhu et al. 2018; Wang et al. 2018b). Moreover, where changes in expression and fitness are situated in a network can help inform which genetic changes may or may not be permissible. One can use regulatory networks, protein–protein interaction networks or genome-scale metabolic models to contextualize the stress response. It turns out that with the inclusion of network features such as degree (how many connections does a gene have) or clustering coefficients (how many of a gene's neighbors are neighbors of each other), machine learning models can be used to predict in which genes adaptive mutations are most likely to occur (Zhu et al. 2018). Moreover, sequence conservation and prevalence, which are features that can be extracted from the pangenome, and which describe how "plastic" (or variable) each gene is, improve prediction accuracy (Fig. 4c) (Zhu et al. 2018). While this is a step toward predicting the emergence of resistance before it actually occurs, for instance during treatment, incorporation of pangenome-wide genetic interaction networks will likely even further enhance the predictive power and accuracy.

1.4 Developing New Therapeutics in the Light of the Pangenome

There is an urgent need to develop new strategies to combat resistant pathogens. Both essential genes and genes required for virulence provide attractive targets for the development of new drugs or biologicals (Clatworthy et al. 2007; Juhas et al.


A.
0
250000
500000
750000
1000000
1250000
1500000
1750000
2000000
B.
Mutation frequency
Fitness change
Essential genes
Sequence prevalence
Sequence conservation
Expression change
90'
20'
Adapted gene
250000
C.
ROC for Classification by SVM
AUC with pangenome features: 0.99778
AUC without pangenome features: 0.91021
True positive rate
False positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

2011; Mobegi et al. 2014). Such candidate targets have been identified for many species by combining functional experimental analyses like Tn-Seq, RNA-Seq, or CRISPRi with computational predictive models (Mobegi et al. 2017b). However, it is becoming more and more apparent that a gene identified as essential or required for infection in one specific strain is not necessarily essential or required for growth or infection in a different genetic context (Rancati et al. 2018). As a consequence, pangenome variability must be taken into consideration when developing new therapeutics that work at a species-wide level.

1.4.1 Targeting Essential Genes

A gene is essential if it is indispensable for reproductive success, which in the case of unicellular organisms are those genes that are required for replication (Rancati et al. 2018). A loss-of-function mutation in one of these genes, or a drug that inactivates its function will stop growth. That is why the identification of a pathogen's essentialome (i.e., the set of essential genes in a defined genome or group of genomes) is an attractive approach for the identification of new drug targets.

Currently, Tn-Seq and related techniques are probably the most popular experimental tools used to determine essentialomes (Peng et al. 2017). Genes that lack insertions in saturated transposon libraries selected in rich media, are considered to be highly likely to be essential in any given condition. CRISPRi is another technique that is rapidly gaining popularity for determining gene-essentiality in both prokaryotes and eukaryotes (Peters et al. 2016; Liu et al. 2017; Wang et al. 2018a). However, since many more strains and species exist than can efficiently and rapidly be experimentally screened for their essential genes, in silico predictive models of gene essentiality are receiving increasing interest (Mobegi et al. 2017b; Nigatu et al.

Fig. 4 Prediction of adaptive evolution relies on pangenome features. (**a**) Circular plot of the *S. pneumoniae* chromosome, with all features necessary for accurate prediction of which genes will contribute with adaptive mutations to vancomycin resistance. Importantly, there is no clear association with any dataset alone and the adaptive outcome, however, when taken as a whole, all data types contribute to distinguishing adapted genes from non-adapted ones (see (**c**) and Zhu et al. 2018). (**b**) Legend for (**a**). From innermost plot to outermost: Expression change: $\log_2$ Fold Change in gene expression comparing vancomycin treatment to no antibiotic treatment after 20, 30, 45, 60, and 90 min of antibiotic exposure. Sequence conservation: $-\log_{10}$ Smith–Waterman distance across all pairs of homologous sequences. Sequence prevalence: percentage of strains in the *S. pneumoniae* pangenome that have a homolog of the gene. Essential genes: genes necessary for survival, as determined by Tn-Seq. Fitness change: change in fitness comparing vancomycin treatment to a no antibiotic control as determined by Tn-Seq. Mutation frequency: frequency of each mutation in a population adapted to vancomycin. Adapted gene: gene containing at least one mutation that is fixed at high frequency, and is specific to the vancomycin adapted populations. (**c**) Classification of adapted genes and non-adapted genes. Receiver operating characteristic curve for a support vector machine trained with all data from (**a**, **b**) (blue) and one trained with pangenome sequence conservation and sequence prevalence omitted (red). The inclusion of these two pangenomic features improves the performance of the classifier

2017). Such models can be based on several types of data, including those obtained from the genomic sequence of an organism (codon usage, orthology, GC content, etc.) or from experimental data such as expression profiles or network topology (Mobegi et al. 2017b; Nigatu et al. 2017). The accuracy of the latter models relies on omics data obtained from species where functional genomics experiments could be performed whereas models based on sequencing are more suitable for poorly studied organisms. Importantly, it is becoming clear that the static concept of gene essentiality is no longer valid. Instead, essentiality is a context-dependent attribute affected by both the environment and the genetic background of a bacterium (Rancati et al. 2018). In the simplest case, a gene can be conditionally essential, meaning it is essential in a specific environment but not in another, or in the case of a pathogen, a gene can be essential in a specific body compartment but not in another.

To understand how genetic context affects gene essentiality it is important to consider the network structure of a genome. Genes in a genome do not act as isolated units, but they interact with each other forming a network. The connections that shape this network can represent protein–protein interactions, epistatic relationships, or transcriptional regulatory interactions (Babu and Madan Babu 2008; Wuchty and Uetz 2014; Costanzo et al. 2016). Some genes present a high degree, i.e. a high number of interactions connecting them to other genes, while other genes are poorly connected (low degree). Essential genes have been shown to have a higher degree than nonessential ones in these genetic interaction networks (Jeong et al. 2001; Davierwala et al. 2005; Costanzo et al. 2010, 2016; Kim et al. 2012; Jiang et al. 2015), which is a characteristic that has been used to predict gene essentiality (Shim et al. 2017). Interestingly, data from different yeast strains has shown that essential genes may be split up into those that are always essential (their loss cannot be overcome), and those that are essential depending on the genetic background. The loss of essential genes from this latter category can be compensated by the adaptive evolution of alternative cellular processes; such essential genes are thereby referred to as “evolvable” essential genes (Motter et al. 2008; Liu et al. 2015). As an example that this is not limited to yeast, the proteins MreC and MreD, involved in peripheral peptidoglycan synthesis, are essential for some *S. pneumoniae* strains. However, different mutations, including the inactivation of the *pbp1a* gene can suppress the essentiality of these proteins (Land and Winkler 2011), which classify *mreC* and *mreD* as “evolvable” essential genes. This evolvability thus at least partially explains how the essentiality of genes can depend on the genetic background and underscores that it is important to determine a pathogens essentialome at a species-wide level to enable the identification of pangenome-wide drug targets. In general, broad-spectrum antibiotics work against large groups of different species of bacteria, and thus existing drugs often target the “pangenome.” Interestingly, these new pangenome concepts are creating opportunities to develop drugs that are directed at a specific clade. The mevalonate pathway is an example of an essential function against which clade-targeting drugs have been developed. This pathway is involved in the production of isoprenoids, and has been shown to be essential in different Gram-positive bacteria (Wilding et al. 2000; Balibar et al. 2009). The pathway is inhibited by an intermediate product, diphosphomevalonate, and fluorinated derivatives of this compound have shown potent antibacterial activities (Kang et al. 2015). However, while the mevalonate pathway is essential, it has also been shown to be evolvable in *S. aureus* (Reichert et al. 2018) raising the possibility that resistance mechanisms can easily arise. Importantly, genomic comparison of different *Staphylococci* has shown that species either have the mevalonate or the non-mevalonate (or 2C-methyl-D-erythritol-4-phosphate, MEP) pathway for the biosynthesis of isoprenoids, and specific pathogenic *Staphylococci* of domestic animals have the non-mevalonate pathway (Misic et al. 2016). Based on this difference at the genus level, it has been proposed that antibiotics for domestic animal *Staphylococci* targeting the MEP pathway could avoid the emergence of antibiotic-resistant determinants in human pathogens. Such clade targeting antibiotics may thus be an interesting strategy, but are only possible if a comprehensive understanding of the pangenome is available.

1.4.2 Targeting Mechanisms of Infection

In addition to genes essential for general growth, genes required for colonization, infection and/or those that damage the host (i.e., virulence factors) are also attractive targets for drug therapies (Clatworthy et al. 2007; Rasko and Sperandio 2010; Allen et al. 2014; Dickey et al. 2017). Consequently, resistance mechanisms against compounds targeting these factors (antivirulence drugs) may not easily spread outside the host (Allen et al. 2014). Also, antivirulence drugs may be more effective against persisters (Kim et al. 2018), and since they are directed at very specific targets they could potentially have less of an effect on the natural microbiota of the host (Clatworthy et al. 2007; Dickey et al. 2017). Specific antivirulence drugs or biologicals at different stages of clinical development, target pathways including the production of teichoic acids, biofilm formation, quorum-sensing mechanisms, and specific histidine kinases, and are directed against bacteria including the ESKAPE pathogens (Matano et al. 2016; Pasquina et al. 2016; Goswami et al. 2017; Dickey et al. 2017; Cardona et al. 2018; Huggins et al. 2018). To expand such specific therapeutic options, it is necessary to identify a pathogen's genetic requirements for infection, for which in vivo Tn-Seq experiments have proven successful (van Opijnen and Camilli 2012; de Vries et al. 2017; Le Breton et al. 2017; Shields et al. 2018). As with essential genes, requirements for certain genes seem to be environment dependent. For example, proline biosynthetic genes in *S. pneumoniae* strain TIGR4 have been shown to be required for infecting mouse lungs, but are dispensable for colonizing the nasopharynx (van Opijnen and Camilli 2012). Other environmental factors that affect genetic requirements are microbial communities and polymicrobial infections. For instance, *S. aureus* requires 182 genes for a successful infection when co-inoculated with *P. aeruginosa*, but the same genes are dispensable if the pathogen is inoculated by itself (Ibberson et al. 2017). By using a Tn-Seq approach, it was shown that two different strains of *P. aeruginosa* required different genes to grow in cystic fibrosis sputum, a growth condition that partially mimics an in vivo infection (Turner et al. 2015). Moreover, many of the genes

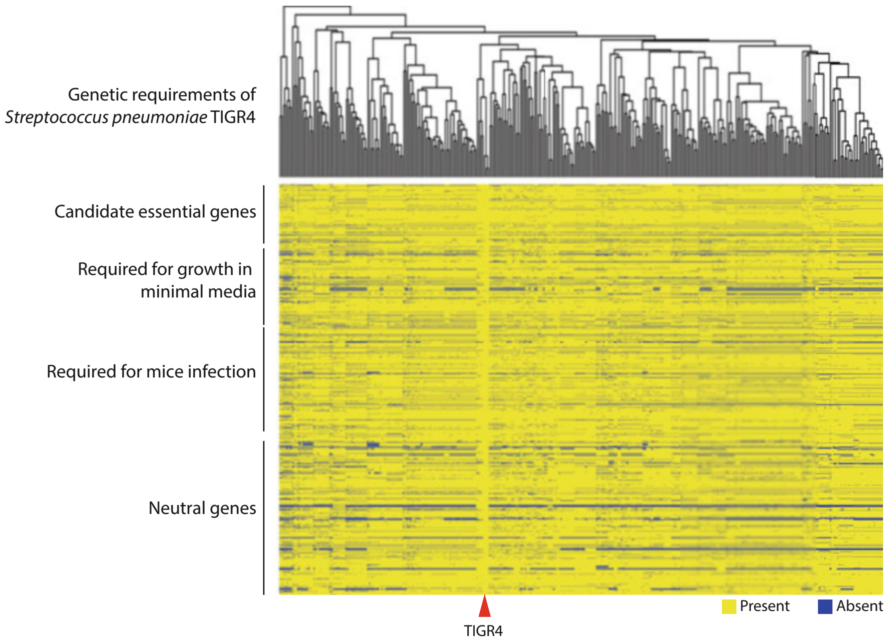


Fig. 5 Genetic requirements of strain *S. pneumoniae*. TIGR4 and its comparison with the species pangenome. Tn-Seq experiments performed in strain TIGR4 (red arrowhead) determined candidate essential genes, genes required for growth in minimal medium and those required for infection (van Opijnen and Camilli 2012). The presence (yellow) and absence (blue) of these genes was established in 332 other *S. pneumoniae* strains. It is clearly shown that many genes identified as essential or required for infection in TIGR4 are absent in many other invasive disease isolates (Cremers et al. 2015)

required by *S. pneumoniae* strain TIGR4 for host colonization are not present in the genomes of other clinical isolates of the species (Fig. 5), which underscores that virulence determinants are indeed also dependent on genetic background and thus only make sense in the context of the pangenome. A successful example of consideration of the pangenome to develop an antibacterial therapy is the pneumococcal vaccine (Berical et al. 2016; Brooks and Mias 2018). The *S. pneumoniae* capsule is one of its most important virulence factors and its diversity is high, with over 90 types (serotypes) currently described (Geno et al. 2015). Capsules are highly antigenic and serotypes differ in polysaccharide residue composition, chemical decoration of sugar monomers and length of the polysaccharide chain (Bentley et al. 2006). Pneumococcal vaccines are formulated by mixing multiple capsule serotypes, which is exemplified by the pneumococcal conjugate vaccine 13 (PCV13) and the pneumococcal polysaccharide vaccine (PPSV23). These vaccines protect against 13 and 23 different capsule-type-based strains, respectively (Berical et al. 2016; Brooks and Mias 2018), and are thereby highly successful in targeting a considerable part of the *S. pneumoniae* pangenome.

1.4.3 Antimicrobial Combination Therapy

In addition to developing novel therapies, utilizing currently available drugs in a more effective way, e.g., through a multidrug strategy (including antibiotic cycling and antimicrobial combination therapy), potentially provides enhanced ways to treat clinical infections and prevent resistance (Smirnova et al. 2011; Yoshida et al. 2017; Firsov et al. 2017). However, it has been shown that the responses to drug–drug combinations can be species specific (even among phylogenetically related organisms), and in some cases strain specific (Brochado et al. 2018). Thus, the application of combination therapies presents a challenge with respect to the pangenome. To overcome this challenge, it is necessary to get a comprehensive understanding of drug–drug interaction outcomes in many species and strains of a species, potentially by testing all possible combinations of drugs, and at different concentrations. Brochado et al. performed 2883 pairwise drug–drug combinations on six bacterial strains from three Gram-negative bacterial species (*E. coli*, *S. typhimurium* and *P. aeruginosa*), yielding a total of 17,050 combinations. The authors found that 70% of the detected drug–drug interactions are species specific, and that 13–30% are strain specific, with different interaction outcomes among the strains (Brochado et al. 2018). Although approaches like these are very important, they can be very time consuming and expensive to perform, especially when one considers hundreds of species/strains in a pangenome. This has prompted the application of computational predictive strategies. One such an approach is INDIGO through which the developers were able to identify a group of genes that are predictive of antibiotic interactions in *E. coli*, and use these genes to predict drug interaction outcomes in other important pathogens including M. tuberculosis and S. aureus (Chandrasekaran et al. 2016). Using such predictive modeling methods considerably reduces the number of experiments that need to be performed, potentially making it possible to accurately infer drug–drug interaction outcomes on a pangenome scale.

2 Conclusions

The development of antibiotic resistance is a complex process that can involve multiple modes of adaptation and/or multiple sources of selective pressure. Here we argue that a fuller understanding of this process is only possible by viewing it through the lens of the pangenome. We have highlighted recent work that demonstrates that genetic background plays an important role in how bacteria respond to an antibiotic and how they develop resistance. We have explained how species and strains with different genetic backgrounds may exhibit (slightly) different adaptational outcomes in response to antibiotics. Strains within a pangenome may also exhibit strain-specific differences in their mechanism and level of resistance as well as their ability to evolve resistance. These different outcomes can be put into context and partially explained by how antibiotic stress is experienced and processed in strain- and species-specific ways. In this way, antibiotics can contribute to the maintenance and shaping of a pangenome by driving adaptive evolution in strain-specific ways.

In addition to providing context for understanding strain- and species-specific responses to antibiotics and their development of resistance, the pangenome can provide a means of predicting the development of resistance as well as inform the development of novel therapeutics. We argue that adaptation to sustained antibiotic pressure is not a wholly stochastic process but rather constrained by a strain's genetic background as well as its environmental context. Given these constraints, it is increasingly possible to utilize machine learning algorithms to make predictions on the probability that a bacterium will evolve resistance. These algorithms can utilize multiple layers of data including genomic, transcriptional, and metabolic datasets at the pangenome level. Therefore, they will continue to improve as additional datasets are generated. Finally, we have considered the role the pangenome could play in developing new therapeutics to combat resistant pathogens. Essential genes and virulence genes offer attractive targets for developing novel therapeutics; however, these targets must be considered within the context of the pangenome due to a variety of reasons. Essential genes in one strain may, in fact, be evolvable, while they are static in another strain. Virulence targets may also be strain specific or dependent on the environmental context of infection. While this may limit the number of targets that are present throughout the pangenome, it does offer the possibility of identifying targets that are strain or species specific.

References

Alam MT, Petit RA 3rd, Crispell EK et al (2014) Dissecting vancomycin-intermediate resistance in *Staphylococcus aureus* using genome-wide association. Genome Biol Evol 6:1174–1185

Albert TJ, Dailidiene D, Dailide G et al (2005) Mutation discovery in bacterial genomes: metronidazole resistance in *Helicobacter pylori*. Nat Methods 2:951–953

Allen RC, Popat R, Diggle SP, Brown SP (2014) Targeting virulence: can we make evolution-proof drugs? Nat Rev Microbiol 12:300–308

Altenhoff AM, Glover NM, Train CM et al (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res 46:D477–D485

Andersson DI, Hughes D (2010) Antibiotic resistance and its cost: is it possible to reverse resistance? Nat Rev Microbiol 8:260–271

Angst DC, Hall AR (2013) The cost of antibiotic resistance depends on evolutionary history in *Escherichia coli*. BMC Evol Biol 13:163

Baba T, Ara T, Hasegawa M et al (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2(2006):0008

Babu MM, Madan Babu M (2008) Computational approaches to study transcriptional regulation. Biochem Soc Trans 36:758–765

Balibar CJ, Shen X, Tao J (2009) The mevalonate pathway of *Staphylococcus aureus*. J Bacteriol 191:851–861

Baquero F, Coque TM, de la Cruz F (2011) Ecology and evolution as targets: the need for novel eco-evo drugs and strategies to fight antibiotic resistance. Antimicrob Agents Chemother 55:3649–3660

Bentley SD, Aanensen DM, Mavroidi A et al (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet 2:e31

Berical AC, Harris D, Dela Cruz CS, Possick JD (2016) Pneumococcal vaccination strategies. An update and perspective. Ann Am Thorac Soc 13:933–944

Blair JMA, Webber MA, Baylay AJ et al (2015) Molecular mechanisms of antibiotic resistance. Nat Rev Microbiol 13:42–51

Blount ZD, Barrick JE, Davidson CJ, Lenski RE (2012) Genomic analysis of a key innovation in an experimental *Escherichia coli* population. Nature 489:513–518

Boissy R, Ahmed A, Janto B et al (2011) Comparative supragenomic analyses among the pathogens *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Haemophilus influenzae* using a modification of the finite supragenome model. BMC Genomics 12:187

Breidenstein EB, Khaira BK, Wiegand I et al (2008) Complex ciprofloxacin resistome revealed by screening a *Pseudomonas aeruginosa* mutant library for altered susceptibility. Antimicrob Agents Chemother 52:4486–4491

Bresler G (2014) Efficiently learning Ising models on arbitrary graphs. arXiv:14116156 [cs, math, stat]

Břinda K, Callendrello A, Cowley L et al (2018) Lineage calling can identify antibiotic resistant clones within minutes. bioRxiv. https://doi.org/10.1101/403204

Brochado AR, Telzerow A, Bobonis J et al (2018) Species-specific activity of antibacterial drug combinations. Nature 559:259–263

Brooks LRK, Mias GI (2018) *Streptococcus pneumoniae*'s virulence and host immunity: aging, diagnostics, and prevention. Front Immunol 9:1366. https://doi.org/10.3389/fimmu.2018.01366

Cardona ST, Choy M, Hogan AM (2018) Essential two-component systems regulating cell envelope functions: opportunities for novel antibiotic therapies. J Membr Biol 251:75–89

Carey AF, Rock JM, Krieger IV et al (2018) TnSeq of *Mycobacterium tuberculosis* clinical isolates reveals strain-specific antibiotic liabilities. PLoS Pathog 14:e1006939

Chancey ST, Agrawal S, Schroeder MR et al (2015) Composite mobile genetic elements disseminating macrolide resistance in *Streptococcus pneumoniae*. Front Microbiol 6:26

Chandrasekaran S, Cokol-Cakmak M, Sahin N et al (2016) Chemogenomics and orthology-based design of antibiotic combination therapies. Mol Syst Biol 12:872

Chewapreecha C, Harris SR, Croucher NJ et al (2014) Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet 46:305–309

Clatworthy AE, Pierson E, Hung DT (2007) Targeting virulence: a new paradigm for antimicrobial therapy. Nat Chem Biol 3:541–548

Comas I, Chakravartti J, Small PM et al (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat Genet 42:498–503

Cornick JE, Bentley SD (2012) *Streptococcus pneumoniae*: the evolution of antimicrobial resistance to beta-lactams, fluoroquinolones and macrolides. Microbes Infect 14:573–583

Costanzo M, Baryshnikova A, Bellay J et al (2010) The genetic landscape of a cell. Science 327:425–431

Costanzo M, VanderSluis B, Koch EN et al (2016) A global genetic interaction network maps a wiring diagram of cellular function. Science 353. https://doi.org/10.1126/science.aaf1420

Cremers AJH, Mobegi FM, de Jonge MI et al (2015) The post-vaccine microevolution of invasive *Streptococcus pneumoniae*. Sci Rep 5:14952

Croucher NJ, Harris SR, Fraser C et al (2011) Rapid pneumococcal evolution in response to clinical interventions. Science 331:430–434

D'Costa VM, McGrann KM, Hughes DW, Wright GD (2006) Sampling the antibiotic resistome. Science 311:374–377

Davierwala AP, Haynes J, Li Z et al (2005) The synthetic genetic interaction spectrum of essential genes. Nat Genet 37:1147–1152

Davies J, Davies D (2010) Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev 74:417–433

de Vries SP, Gupta S, Baig A et al (2017) Genome-wide fitness analyses of the foodborne pathogen *Campylobacter jejuni* in in vitro and in vivo models. Sci Rep 7:1251

Del Barrio-Tofiño E, López-Causapé C, Cabot G, et al (2017) Genomics and susceptibility profiles of extensively drug-resistant *Pseudomonas aeruginosa* isolates from Spain. Antimicrob Agents Chemother 61. doi: https://doi.org/10.1128/AAC.01589-17

Desjardins CA, Cohen KA, Munsamy V et al (2016) Genomic and functional analyses of *Mycobacterium tuberculosis* strains implicate ald in D-cycloserine resistance. Nat Genet 48:544–551

Dickey SW, Cheung GYC, Otto M (2017) Different drugs for bad bugs: antivirulence strategies in the age of antibiotic resistance. Nat Rev Drug Discov 16:457–471

Drlica K, Malik M, Kerns RJ, Zhao X (2008) Quinolone-mediated bacterial death. Antimicrob Agents Chemother 52:385–392

El'Garch F, Jeannot K, Hocquet D et al (2007) Cumulative effects of several nonenzymatic mechanisms on the resistance of *Pseudomonas aeruginosa* to aminoglycosides. Antimicrob Agents Chemother 51:1016–1021

Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet 4:457–469

Fàbrega A, Madurga S, Giralt E, Vila J (2009) Mechanism of action of and resistance to quinolones. Microb Biotechnol 2:40–61

Fajardo A, Martinez-Martin N, Mercadillo M et al (2008) The neglected intrinsic resistome of bacterial pathogens. PLoS One 3:e1619

Farhat MR, Shapiro BJ, Kieser KJ et al (2013) Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet 45:1183–1189

Fenner L, Egger M, Bodmer T et al (2012) Effect of mutation and genetic background on drug resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother 56:3047–3053

Firsov AA, Golikova MV, Strukova EN et al (2017) Pharmacokinetically-based prediction of the effects of antibiotic combinations on resistant *Staphylococcus aureus* mutants: in vitro model studies with linezolid and rifampicin. J Chemother 29:220–226

Floss HG, Yu TW (2005) Rifamycin-mode of action, resistance, and biosynthesis. Chem Rev 105:621–632

Frank KL, Bundle SF, Kresge ME et al (2003) aadA confers streptomycin resistance in Borrelia burgdorferi. J Bacteriol 185:6723–6727

Gagneux S, Burgos MV, DeRiemer K et al (2006) Impact of bacterial genetics on the transmission of isoniazid-resistant *Mycobacterium tuberculosis*. PLoS Pathog 2:e61

Gallagher LA, Shendure J, Manoil C (2011) Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq. MBio 2:e00315–e00310

Gawronski JD, Wong SM, Giannoukos G et al (2009) Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. Proc Natl Acad Sci USA 106:16422–16427

Geisinger E, Vargas-Cuebas G, Mortman NJ, Syal S, Dai Y, Wainwright EL, Lazinski D, Wood S, Zhu Z, Anthony J, van Opijnen T, Isberg RR (2019) The landscape of phenotypic and transcriptional responses to ciprofloxacin in : acquired resistance alleles modulate drug-induced SOS response and Prophage replication. MBio 10(3)

Geno KA, Gilbert GL, Song JY et al (2015) Pneumococcal capsules and their types: past, present, and future. Clin Microbiol Rev 28:871–899

Gifford DR, Furio V, Papkou A et al (2018) Identifying and exploiting genes that potentiate the evolution of antibiotic resistance. Nat Ecol Evol 2:1033–1039

Girgis HS, Hottes AK, Tavazoie S (2009) Genetic architecture of intrinsic antibiotic susceptibility. PLoS One 4:e5629

Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, Lozupone CA, Knight R, Gordon JI (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe 6(3):279–289

Goswami M, Wilke KE, Carlson EE (2017) Rational design of selective adenine-based scaffolds for inactivation of bacterial histidine kinases. J Med Chem 60:8170–8182

Gullberg E, Cao S, Berg OG et al (2011) Selection of resistant bacteria at very low antibiotic concentrations. PLoS Pathog 7:e1002158

Gygli SM, Borrell S, Trauner A, Gagneux S (2017) Antimicrobial resistance in *Mycobacterium tuberculosis*: mechanistic and evolutionary perspectives. FEMS Microbiol Rev 41:354–373

Hall AR, MacLean RC (2016) Epistasis buffers the fitness effects of rifampicin-resistance mutations in *Pseudomonas aeruginosa*. Evolution 70:1161–1161

Hershberg R, Lipatov M, Small PM et al (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol 6:e311

Huggins WM, Barker WT, Baker JT et al (2018) Meridianin D analogues display antibiofilm activity against MRSA and increase Colistin efficacy in gram-negative bacteria. ACS Med Chem Lett 9:702–707

Ibberson CB, Stacy A, Fleming D et al (2017) Co-infecting microorganisms dramatically alter pathogen gene essentiality during polymicrobial infection. Nat Microbiol 2:17079

Ising E (1925) Beitrag zur Theorie des Ferromagnetismus. Z Phys 31:253–258

Jacobs MA, Alwood A, Thaipisuttikul I et al (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. Proc Natl Acad Sci USA 100:14339–14344

Jensen PA, Zhu Z, van Opijnen T (2017) Antibiotics disrupt coordination between transcriptional and phenotypic stress responses in pathogenic bacteria. Cell Rep 20:1705–1716

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42

Jiang P, Wang H, Li W et al (2015) Network analysis of gene essentiality in functional genomics experiments. Genome Biol 16:239

Juhas M, Eberl L, Glass JI (2011) Essence of life: essential genes of minimal genomes. Trends Cell Biol 21:562–568

Kang S, Watanabe M, Jacobs JC et al (2015) Synthesis of mevalonate- and fluorinated mevalonate prodrugs and their in vitro human plasma stability. Eur J Med Chem 90:448–461

Kauffman SA, Weinberger ED (1989) The NK model of rugged fitness landscapes and its application to maturation of the immune response. J Theor Biol 141:211–245

Kim J, Kim I, Han SK et al (2012) Network rewiring is an important mechanism of gene essentiality change. Sci Rep 2:900

Kim W, Hendricks GL, Tori K et al (2018) Strategies against methicillin-resistant *Staphylococcus aureus* persisters. Future Med Chem 10:779–794

Kohanski MA, Collins JJ (2008) Rewiring bacteria, two components at a time. Cell 133:947–948

Kohanski MA, Dwyer DJ, Collins JJ (2010) How antibiotics kill bacteria: from targets to networks. Nat Rev Microbiol 8:423–435

Koskiniemi S, Pranting M, Gullberg E et al (2011) Activation of cryptic aminoglycoside resistance in *Salmonella enterica*. Mol Microbiol 80:1464–1478

Lakin SM, Dean C, Noyes NR et al (2017) MEGARes: an antimicrobial resistance database for high throughput sequencing. Nucleic Acids Res 45:D574–D580

Lampson BC, von David W, Parisi JT (1986) Novel mechanism for plasmid-mediated erythromycin resistance by pNE24 from *Staphylococcus epidermidis*. Antimicrob Agents Chemother 30:653–658

Land AD, Winkler ME (2011) The requirement for pneumococcal MreC and MreD is relieved by inactivation of the gene encoding PBP1a. J Bacteriol 193:4166–4179

Lang GI, Rice DP, Hickman MJ et al (2013) Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature 500:571–574

Langridge GC, Phan MD, Turner DJ et al (2009) Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res 19:2308–2316

Le Breton Y, Belew AT, Freiberg JA et al (2017) Genome-wide discovery of novel M1T1 group A streptococcal determinants important for fitness and virulence during soft-tissue infection. PLoS Pathog 13:e1006584

Lee AS, Teo AS, Wong SY (2001) Novel mutations in ndh in isoniazid-resistant *Mycobacterium tuberculosis* isolates. Antimicrob Agents Chemother 45:2157–2159

Lind PA, Farr AD, Rainey PB (2015) Experimental evolution reveals hidden diversity in evolutionary pathways. Elife 4. https://doi.org/10.7554/eLife.07074

Lindgren PK, Karlsson Å, Hughes D (2003) Mutation rate and evolution of fluoroquinolone resistance in *Escherichia coli* isolates from patients with urinary tract infections. Antimicrob Agents Chemother 47:3222–3232

Liu F, Zhu Y, Yi Y et al (2014) Comparative genomic analysis of *Acinetobacter baumannii* clinical isolates reveals extensive genomic variation and diverse antibiotic resistance determinants. BMC Genomics 15:1163

Liu G, Yong MY, Yurieva M et al (2015) Gene essentiality is a quantitative property linked to cellular evolvability. Cell 163:1388–1399

Liu X, Gallay C, Kjos M et al (2017) High-throughput CRISPRi phenotyping identifies new essential genes in *Streptococcus pneumoniae*. Mol Syst Biol 13:931

Ma W, Zhang D, Li G et al (2017) Antibacterial mechanism of daptomycin antibiotic against *Staphylococcus aureus* based on a quantitative bacterial proteome analysis. J Proteome 150:242–251

Magrini V, Creighton C, White D et al (1998) The aadA gene of plasmid R100 confers resistance to spectinomycin and streptomycin in *Myxococcus xanthus*. J Bacteriol 180:6757–6760

Mark BL, Vocadlo DJ, Oliver A (2011) Providing beta-lactams a helping hand: targeting the AmpC beta-lactamase induction pathway. Future Microbiol 6:1415–1427

Matano LM, Morris HG, Wood BM et al (2016) Accelerating the discovery of antibacterial compounds using pathway-directed whole cell screening. Bioorg Med Chem 24:6307–6314

Matejkova P, Flasarova M, Zakoucka H et al (2009) Macrolide treatment failure in a case of secondary syphilis: a novel A2059G mutation in the 23S rRNA gene of Treponema pallidum subsp. pallidum. J Med Microbiol 58:832–836

McArthur AG, Waglechner N, Nizam F et al (2013) The comprehensive antibiotic resistance database. Antimicrob Agents Chemother 57:3348–3357

McDermott PF, Tyson GH, Kabera C et al (2016) Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal Salmonella. Antimicrob Agents Chemother 60:5515–5520

McKeegan KS, Borges-Walmsley MI, Walmsley AR (2002) Microbial and viral drug resistance mechanisms. Trends Microbiol 10:S8–S14

Meier A, Kirschner P, Bange FC et al (1994) Genetic alterations in streptomycin-resistant *Mycobacterium tuberculosis*: mapping of mutations conferring resistance. Antimicrob Agents Chemother 38:228–233

Melnyk AH, Wong A, Kassen R (2015) The fitness costs of antibiotic resistance mutations. Evol Appl 8:273–283

Misic AM, Cain CL, Morris DO et al (2016) Divergent isoprenoid biosynthesis pathways in *Staphylococcus* species constitute a drug target for treating infections in companion animals. mSphere 1. https://doi.org/10.1128/mSphere.00258-16

Mobegi FM, van Hijum SA, Burghout P et al (2014) From microbial gene essentiality to novel antimicrobial drug targets. BMC Genomics 15:958

Mobegi FM, Cremers AJH, de Jonge MI et al (2017a) Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data. Sci Rep 7:42808

Mobegi FM, Zomer A, de Jonge MI, van Hijum SA (2017b) Advances and perspectives in computational prediction of microbial gene essentiality. Brief Funct Genomics 16:70–79

Motter AE, Gulbahce N, Almaas E, Barabasi AL (2008) Predicting synthetic rescues in metabolic networks. Mol Syst Biol 4:168

Moya B, Dotsch A, Juan C et al (2009) Beta-lactam resistance response triggered by inactivation of a nonessential penicillin-binding protein. PLoS Pathog 5:e1000353

Munita JM, Arias CA (2016) Mechanisms of antibiotic resistance. Microbiol Spectr 4. https://doi.org/10.1128/microbiolspec.VMBF-0016-2015

Murakami K, Nomura K, Doi M, Yoshida T (1987) Production of low-affinity penicillin-binding protein by low- and high-resistance groups of methicillin-resistant *Staphylococcus aureus*. Antimicrob Agents Chemother 31:1307–1311

Murray JL, Kwon T, Marcotte EM, Whiteley M (2015) Intrinsic antimicrobial resistance determinants in the superbug *Pseudomonas aeruginosa*. MBio 6:e01603–e01615

Nash KA, Inderlied CB (1995) Genetic basis of macrolide resistance in *Mycobacterium avium* isolated from patients with disseminated disease. Antimicrob Agents Chemother 39:2625–2630

Nichols RJ, Sen S, Choo YJ et al (2011) Phenotypic landscape of a bacterial cell. Cell 144:143–156

Nigatu D, Sobetzko P, Yousef M, Henkel W (2017) Sequence-based information-theoretic features for gene essentiality prediction. BMC Bioinformatics 18:473

Okamoto S, Tamaru A, Nakajima C et al (2007) Loss of a conserved 7-methylguanosine modification in 16S rRNA confers low-level streptomycin resistance in bacteria. Mol Microbiol 63:1096–1106

Pasquina L, Santa Maria JP Jr, McKay Wood B et al (2016) A synthetic lethal approach for compound and target identification in *Staphylococcus aureus*. Nat Chem Biol 12:40–45

Peng C, Lin Y, Luo H, Gao F (2017) A comprehensive overview of online resources to identify and predict bacterial essential genes. Front Microbiol 8:2331

Perdigao J, Macedo R, Machado D et al (2014) GidB mutation as a phylogenetic marker for Q1 cluster *Mycobacterium tuberculosis* isolates and intermediate-level streptomycin resistance determinant in Lisbon, Portugal. Clin Microbiol Infect 20:O278–O284

Pérez-Llarena FJ, Bou G (2016) Proteomics as a tool for studying bacterial virulence and antimicrobial resistance. Front Microbiol 7:410

Peters JM, Colavin A, Shi H et al (2016) A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. Cell 165:1493–1506

Portevin D, Gagneux S, Comas I, Young D (2011) Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. PLoS Pathog 7:e1001307

Power RA, Parkhill J, de Oliveira T (2017) Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet 18:41–50

Price MN, Wetmore KM, Waters RJ et al (2018) Mutant phenotypes for thousands of bacterial genes of unknown function. Nature 557:503–509

Prudhomme M, Attaiech L, Sanchez G et al (2006) Antibiotic stress induces genetic transformability in the human pathogen *Streptococcus pneumoniae*. Science 313:89–92

Qin H, Lo NW-S, Loo JF-C et al (2018) Comparative transcriptomics of multidrug-resistant *Acinetobacter baumannii* in response to antibiotic treatments. Sci Rep 8:3515

Quémard A, Lacave C, Lanéelle G (1991) Isoniazid inhibition of mycolic acid synthesis by cell extracts of sensitive and resistant strains of *Mycobacterium aurum*. Antimicrob Agents Chemother 35:1035–1039

Rancati G, Moffat J, Typas A, Pavelka N (2018) Emerging and evolving concepts in gene essentiality. Nat Rev Genet 19:34–49

Rasko DA, Sperandio V (2010) Anti-virulence strategies to combat bacteria-mediated disease. Nat Rev Drug Discov 9:117–128

Reichert S, Ebner P, Bonetti EJ et al (2018) Genetic adaptation of a mevalonate pathway deficient mutant in *Staphylococcus aureus*. Front Microbiol 9:1539

Reichmann P, Konig A, Marton A, Hakenbeck R (1996) Penicillin-binding proteins as resistance determinants in clinical isolates of *Streptococcus pneumoniae*. Microb Drug Resist 2:177–181

Roguev A, Bandyopadhyay S, Zofall M et al (2008) Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. Science 322:405–410

Ropy A, Cabot G, Sanchez-Diener I et al (2015) Role of *Pseudomonas aeruginosa* low-molecular-mass penicillin-binding proteins in AmpC expression, beta-lactam resistance, and peptidoglycan structure. Antimicrob Agents Chemother 59:3925–3934

Rosdahl VT (1985) Localisation of the penicillinase gene in naturally occurring *Staphylococcus aureus* strains. Acta Pathol Microbiol Immunol Scand B 93:383–388

Rybak MJ (2006) Pharmacodynamics: relation to antimicrobial resistance. Am J Infect Control 34: S38–S45; discussion S64–73

Salverda MLM, Dellus E, Gorter FA et al (2011) Initial mutations direct alternative pathways of protein evolution. PLoS Genet 7:e1001321

San Millan A, Heilbron K, MacLean RC (2014) Positive epistasis between co-infecting plasmids promotes plasmid survival in bacterial populations. ISME J 8:601–612

Sandgren A, Strong M, Muthukrishnan P et al (2009) Tuberculosis drug resistance mutation database. PLoS Med 6:e2

Santajit S, Indrawattana N (2016) Mechanisms of antimicrobial resistance in ESKAPE pathogens. Biomed Res Int 2016:2475067

Sauvage E, Kerff F, Fonze E et al (2002) The 2.4-a crystal structure of the penicillin-resistant penicillin-binding protein PBP5fm from *Enterococcus faecium* in complex with benzylpenicillin. Cell Mol Life Sci 59:1223–1232

Schroeder MR, Stephens DS (2016) Macrolide resistance in *Streptococcus pneumoniae*. Front Cell Infect Microbiol 6:98

Schurek KN, Marr AK, Taylor PK et al (2008) Novel genetic determinants of low-level aminoglycoside resistance in *Pseudomonas aeruginosa*. Antimicrob Agents Chemother 52:4213–4219

Shields RC, Zeng L, Culp DJ, Burne RA (2018) Genomewide identification of essential genes and fitness determinants of *Streptococcus mutans* UA159. mSphere 3. https://doi.org/10.1128/mSphere.00031-18

Shim JE, Lee T, Lee I (2017) From sequencing data to gene functions: co-functional network approaches. Anim Cells Syst 21:77–83

Sibold C, Henrichsen J, Konig A et al (1994) Mosaic pbpX genes of major clones of penicillin-resistant *Streptococcus pneumoniae* have evolved from pbpX genes of a penicillin-sensitive Streptococcus oralis. Mol Microbiol 12:1013–1023

Silva RF, Mendonça SCM, Carvalho LM et al (2011) Pervasive sign epistasis between conjugative plasmids and drug-resistance chromosomal mutations. PLoS Genet 7. https://doi.org/10.1371/journal.pgen.1002181

Skov RL, Williams TJ, Pallesen L et al (1995) beta-Lactamase production and genetic location in *Staphylococcus aureus*: introduction of a beta-lactamase plasmid in strains of phage group II. J Hosp Infect 30:111–124

Slager J, Kjos M, Attaiech L, Veening J-W (2014) Antibiotic-induced replication stress triggers bacterial competence by increasing gene dosage near the origin. Cell 157:395–406

Smirnova MV, Strukova EN, Portnoy YA et al (2011) The antistaphylococcal pharmacodynamics of linezolid alone and in combination with doxycycline in an in vitro dynamic model. J Chemother 23:140–144

Sommer MOA, Munck C, Toft-Kehler RV, Andersson DI (2017) Prediction of antibiotic resistance: time for a new preclinical paradigm? Nat Rev Microbiol 15:689–696

Stamm LV, Bergen HL (2000) A point mutation associated with bacterial macrolide resistance is present in both 23S rRNA genes of an erythromycin-resistant *Treponema pallidum* clinical isolate. Antimicrob Agents Chemother 44:806–807

Stoebel DM, Dorman CJ (2010) The effect of mobile element IS10 on experimental regulatory evolution in *Escherichia coli*. Mol Biol Evol 27:2105–2112

Stoesser N, Batty EM, Eyre DW et al (2013) Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. J Antimicrob Chemother 68:2234–2244

Svab Z, Harper EC, Jones JD, Maliga P (1990) Aminoglycoside-3″-adenyltransferase confers resistance to spectinomycin and streptomycin in *Nicotiana tabacum*. Plant Mol Biol 14:197–205

Tait RC, Rempel H, Rodriguez RL, Kado CI (1985) The aminoglycoside-resistance operon of the plasmid pSa: nucleotide sequence of the streptomycin-spectinomycin resistance gene. Gene 36:97–104

Tamae C, Liu A, Kim K et al (2008) Determination of antibiotic hypersensitivity among 4,000 single-gene-knockout mutants of *Escherichia coli*. J Bacteriol 190:5981–5988

Thibault D, Jensen PA, Wood S, Qabar C, Clark S, Shainheit MG, Isberg RR, van Opijnen T (2019) Droplet Tn-Seq combines microfluidics with Tn-Seq for identifying complex single-cell phenotypes. Nat Commun 10(1)

Toh SM, Xiong L, Arias CA et al (2007) Acquisition of a natural resistance gene renders a clinical strain of methicillin-resistant *Staphylococcus aureus* resistant to the synthetic antibiotic linezolid. Mol Microbiol 64:1506–1514

Tomasz A (1979) The mechanism of the irreversible antimicrobial effects of penicillins: how the beta-lactam antibiotics kill and lyse bacteria. Annu Rev Microbiol 33:113–137

Tong AHY, Lesage G, Bader GD et al (2004) Global mapping of the yeast genetic interaction network. Science 303:808–813

Turner KH, Wessel AK, Palmer GC et al (2015) Essential genome of *Pseudomonas aeruginosa* in cystic fibrosis sputum. Proc Natl Acad Sci USA 112:4110–4115

Vakulenko SB, Mobashery S (2003) Versatility of aminoglycosides and prospects for their future. Clin Microbiol Rev 16:430–450

van Opijnen T, Camilli A (2012) A fine scale phenotype-genotype virulence map of a bacterial pathogen. Genome Res 22:2541–2551

van Opijnen T, Bodi KL, Camilli A (2009) Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods 6:767–772

van Opijnen T, Dedrick S, Bento J (2016) Strain dependent genetic networks for antibiotic-sensitivity in a bacterial pathogen with a large pan-genome. PLoS Pathog 12:e1005869

Vilchèze C, Jacobs WR Jr (2014) Resistance to isoniazid and ethionamide in *Mycobacterium tuberculosis*: genes, mutations, and causalities. Microbiol Spectr 2:MGM2-0014-2013

Vogwill T, Kojadinovic M, MacLean RC (2016) Epistasis between antibiotic resistance mutations and genetic background shape the fitness effect of resistance across species of *Pseudomonas*. Proc Biol Sci 283. https://doi.org/10.1098/rspb.2016.0151

Walker TM, Kohl TA, Omar SV et al (2015) Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis 15:1193–1202

Wallace RJ Jr, Meier A, Brown BA et al (1996) Genetic basis for clarithromycin resistance among isolates of *Mycobacterium chelonae* and *Mycobacterium abscessus*. Antimicrob Agents Chemother 40:1676–1681

Walsh C (2000) Molecular mechanisms that confer antibacterial drug resistance. Nature 406:775–781

Wang G, Taylor DE (1998) Site-specific mutations in the 23S rRNA gene of *Helicobacter pylori* confer two types of resistance to macrolide-lincosamide-streptogramin B antibiotics. Antimicrob Agents Chemother 42:1952–1958

Wang T, Guan C, Guo J et al (2018a) Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. Nat Commun 9:2475

Wang X, Zorraquino V, Kim M et al (2018b) Predicting the evolution of *Escherichia coli* by a data-driven approach. Nat Commun 9:3562

Watkinson AJ, Murby EJ, Costanzo SD (2007) Removal of antibiotics in conventional and advanced wastewater treatment: implications for environmental discharge and wastewater recycling. Water Res 41:4164–4176

Wattam AR, Davis JJ, Assaf R et al (2017) Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res 45:D535–D542

Weinreich DM, Delaney NF, DePristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312:111–114

Weldhagen GF (2004) Integrons and beta-lactamases—a novel perspective on resistance. Int J Antimicrob Agents 23:556–562

Wetmore KM, Price MN, Waters RJ et al (2015) Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. MBio 6:e00306–e00315

Wielders CLC, Fluit AC, Brisse S et al (2002) mecA gene is widely disseminated in *Staphylococcus aureus* population. J Clin Microbiol 40:3970–3975

Wilding EI, Brown JR, Bryant AP et al (2000) Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in gram-positive cocci. J Bacteriol 182:4319–4327

Wistrand-Yuen E, Knopp M, Hjort K et al (2018) Evolution of high-level resistance during low-level antibiotic exposure. Nat Commun 9:1599

Wong SY, Lee JS, Kwak HK et al (2011) Mutations in gidB confer low-level streptomycin resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother 55:2515–2522

Wright GD (2003) Mechanisms of resistance to antibiotics. Curr Opin Chem Biol 7:563–569

Wuchty S, Uetz P (2014) Protein-protein interaction networks of *E. coli* and *S. cerevisiae* are similar. Sci Rep 4:7187

Yoshida M, Reyes SG, Tsuda S et al (2017) Time-programmable drug dosing allows the manipulation, suppression and reversal of antibiotic drug resistance in vitro. Nat Commun 8:15589

Zampieri M, Enke T, Chubukov V et al (2017a) Metabolic constraints on the evolution of antibiotic resistance. Mol Syst Biol 13:917

Zampieri M, Zimmermann M, Claassen M, Sauer U (2017b) Nontargeted metabolomics reveals the multilevel response to antibiotic perturbations. Cell Rep 19:1214–1228

Zankari E, Hasman H, Cosentino S et al (2012) Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644

Zeng X, Kwok JS, Yang KY et al (2018) Whole genome sequencing data of 1110 *Mycobacterium tuberculosis* isolates identifies insertions and deletions associated with drug resistance. BMC Genomics 19:365

Zhu Z, Surujon D, Pavao A et al (2018) Forecasting bacterial survival-success and adaptive evolution through multi-omics stress response-mapping, network analyses and machine learning. bioRxiv. https://doi.org/10.1101/387910

Zhu Z, Surujon D, Ortiz-Marquez JC, Wood S, Huo W, Isberg RR, van Opijnen T (2019) Entropy of a bacterial stress response is a generalizable predictor for fitness and antibiotic sensitivity. bioRxiv. https://doi.org/10.1101/813709

Part III
Pangenomics: An Open, Evolving Discipline

Meta-Pangenome: At the Crossroad of Pangenomics and Metagenomics

Bing Ma, Michael France, and Jacques Ravel

Abstract With the recent technological advancement in cultivation-independent high-throughput sequencing, metagenomes have tremendously improved our ability to characterize the genomic contents of the whole microbial communities. In this chapter, we argue the notion of pangenome can be applied beyond the available genome sequences by leveraging metagenome-assembled genomes, to form a comprehensive representation of the genetic content of a taxonomic group in a particular environment. We present the concept of the meta-pangenome, a representation of the totality of genes belonging to a species identified in multiple metagenomic samplings of a particular habitat. As an essential component in genome-centric pangenome analyses, we emphasize the importance to perform stringent quality assessment and validation to ensure the high quality of metagenomic deconvoluted genomes. This expansion from the traditional pangenome concept to the meta-pangenome overcomes many of the biases associated with whole-genome sequencing, and addresses the in vivo ecological context to further develop a systems-level understanding of microbial ecosystems.

Keywords Meta-pangenome · Pan-metagenome · Pangenome · Metagenome · Comparative genomics · Metagenome-assembled genome · Intraspecies diversity · Metagenomic subspecies · Community ecotype · Habitome

1 Introduction

The first microbial genome, *Haemophilus influenzae*, was sequenced in 1995 (Fleischmann et al. 1995) with the second, *Mycoplasma genitalium*, following a few months later (Fraser et al. 1995). In analyzing the *M. genitalium* genome, the authors compared its sequence to that of *H. influenzae*, the only other available

B. Ma (✉) · M. France · J. Ravel
Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD, USA
e-mail: bma@som.umaryland.edu

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_9

genome sequence at the time, providing insights into the ecology and evolution of these two microbes. Every subsequent genome comparison enabled the identification of shared and unique genetic characteristics between sets organisms. From this observation emerged the concept of pangenome, which describes the core (genes present in every strain of the species) and accessory (genes present in a subset of strains) genomes. Studying the similarities and differences between the genomic content of organisms can inform their evolutionary relationships, ecological roles, relationship to health, and has revolutionized our understanding of microbial diversity (Touchman 2010; Xia 2013; Hardison 2003; Miller et al. 2004; France et al. 2016).

Over the years, and with significant technological advancement, the number of available genome sequences has expanded from a few to a seemingly endless catalog. Yet this impressive collection suffers from a rather severe bias toward species and strains that are related to human health, amenable to isolation, and/or generally tractable. Metagenomics, the sequencing of whole microbial communities, is filling in these gaps by characterizing the genomes of entire populations in a community without cultivation. In this chapter, we argue the notion of pangenome can be applied beyond the available genome sequences by leveraging metagenome-assembled genomes (MAGs), to form a comprehensive representation of the genetic content of a taxonomic group in a particular environment. We present the concept of the meta-pangenome, a representation of the totality of genes belonging to a species identified in multiple metagenomic samplings of a particular habitat. This expansion from the traditional pangenome concept to the meta-pangenome overcomes many of the biases associated with whole-genome sequencing and addresses the in vivo ecological context by describing the whole genetic potential of a species in a specific environment. Further building on this new concept, one can think of the pan-metagenome as the complete genes/proteins catalog of all species found in a giving environment.

2 Metagenome Deconvolution Enables Genome-Centric Analyses of Microbial Ecosystems

An overwhelming majority of microbial species have resisted cultivation in the laboratory, largely due to strict, yet unknown, growth requirements (Bakken 1985). The cultivation of fastidious microbes requires optimal combinations of nutrients, growth temperatures, oxygen levels or even, in some cases, and the presence of key microbial partners (Amann et al. 1995; Eckburg et al. 2005). The inability to grow these organisms has undoubtedly limited our understanding of the ecology of indigenous microbial communities. State-of-the-art whole community sequencing technology via metagenomics has opened the door to in vivo studies of microbial populations and communities. By definition, metagenomic sequencing characterizes the collection of all the genetic material isolated from an environmental sample without traditional cultivation (Handelsman 2004; Iverson et al. 2012; Mackelprang et al. 2011). This has aided the development of systems-level insights into the structure and function of microbial ecosystems (Handelsman 2004; Gilbert and Dupont 2011). Advancements in sequencing technologies and throughput have, and continue to improve our ability to characterize the genomic contents of microbial communities down to the rare biosphere (Eckburg et al. 2005; Sogin et al. 2006).

Metagenomic sequencing results in a dataset of sequence reads that belong to the various species that make up the microbial community. Assembly of these datasets into stretches of contiguous DNA sequences, termed contigs, can be complicated by the presence of conserved genomic regions across species. Development of metagenomic specific short reads assembly algorithms and tools that can disentangle these similar sequences originating from different taxa has improved the quality of metagenomic assemblies (Pevzner et al. 2001), those include IDBA-UD (Peng et al. 2012), MetaVelvet (Namiki et al. 2012), SOAPdenovo (Li et al. 2010; Luo et al. 2012), ABYSS (Simpson et al. 2009), Khmer (Pell et al. 2012; Howe et al. 2012), Ray-meta (Boisvert et al. 2012), MEGAHIT (Li et al. 2015, 2016), and metaSPAdes (Nurk et al. 2017). Binning of these contigs based on genomic characteristics like GC content, tetramer frequency, sequence coverage, among others has enabled researchers to identify sets of contigs that belong to the same species. These advancements have resulted in the concept of metagenome-assembled genomes (MAGs), which represent the collection of all contigs or scaffolds from a single or closely related strains of a given species. Developments in bioinformatics tools used in assembly and binning have made the recovery of genomes from metagenomic datasets a routine analysis, including rare species and draft genomes from previously uncultivated species (Albertsen et al. 2013). Binning algorithms and tools have been reviewed previously (Sangwan et al. 2016; Breitwieser et al. 2017). For each species, the genetic contents of all strains in the population are included in a species bin, although sequencing depth, library construction methods, presence of host DNAs, and other factors may affect the metagenomic sequencing results (Zaheer et al. 2018; Pereira-Marques et al. 2019; Bowers et al. 2015).

MAGs have led to the discovery of a remarkable amount of genomic diversity and the characterization of novel bacterial membership. However, MAGs should always be used with caution for the reasons discussed above. False positives in binning, conflicted, and incomplete MAGs have been observed for a variety of different binning tools that can reduce the quality of public genome repositories if MAGs are not evaluated carefully (Shaiber and Eren 2019). Multiple studies have suggested that downstream MAGs quality assessment and validation steps are critical, and available tools published recently to serve such purpose include MetaQUAST (Koren and Phillippy 2015), CheckM (Parks et al. 2015), MAGpy (Stewart et al. 2019), Anvio (Eren et al. 2015), AMBER (Meyer et al. 2018), and DAS tool (Sieber et al. 2018). Further refinement, stringent quality assessment, extending assembly length through re-assembly after recruiting reads back to the MAGs, and genome completeness assessments are important and necessary steps to ensure the fidelity of the MAGs (Eren et al. 2015). High-quality metagenome-deconvoluted genomes are essential to perform genome-centric in vivo analyses of microbial ecosystems.

3 Metagenome-Assembled Genomes Revealed Extensive within Community Intraspecies Diversity in a Microbial Community

Microbial populations often composed of multiple strains of each species, and the resulting intraspecies diversity could have significant functional and clinical implications (Kraal et al. 2014; Greenblum et al. 2015; Oh et al. 2014). Gel microdroplet cultivation afforded nearly finished single genomes and revealed substantial intraspecies diversity within human oral and fecal microbiomes (Fitzsimons et al. 2013). Strains of dominant human skin bacterial species were shown to be heterogeneous and multiphyletic, which the authors suggested to be the result of micro-scale differences in the environment that shaped the ecology and evolution of each subpopulation (Oh et al. 2014). Another study reported extensive strain-level variation detected in the human gut microbiome using large-scale intraspecies copy number variation (Greenblum et al. 2015). This intraspecies variation is thought to be associated with obesity and inflammatory bowel disease. These studies highlight the complex relationships between within-species diversity and functional capacity, linking compositional shifts to subspecies-level variations.

Intra-species diversity obviously complicates MAGs generation, a problem that is compounded by the use of short-read sequencing technology. It is difficult to establish linkage and synteny between genotypes in a species genome. Binning strategies can separate sequences that belong to different species, but are generally not capable of distinguishing between strains of the same species in a metagenomic dataset (Huson et al. 2011). There are encouraging developments in binning algorithms recently that have addressed strain-level resolution from metagenomic short-read sequencing such as StrainPhlAn (Truong et al. 2017), ConStrains (Luo et al. 2015), MetaSNV (Costea et al. 2017), and DESMAN (Quince et al. 2017). However, the word "strain" has been used interchangeably with subspecies type, genotype, biotype, among others, in metagenome-derived strain-level resolution analyses. Although intraspecies diversity can be purged during assembly, the remainder often leads to species bins that contain composite genetic information from multiple genotypes (strains) of the species. Advancements in chromosome conformation capture (Hi-C) and long-read sequencing technologies such as PacBio SMRT sequencing and Oxford nanopore technologies could improve strain deconvolution from metagenomic data by extending the read length and assembly quality (Frank et al. 2016; Tsai et al. 2016; Belton et al. 2012). However, these technologies have not been widely adopted probably due to technical limitations.

4 A Practical Definition of Meta-Pangenome

The pangenome has been an important concept and a tool used in comparative genomics to dissect microbial diversity. A pangenome generally refers to the entire collection of genetic content from all strains of a species (Tettelin et al. 2005; Medini et al. 2005; Vernikos et al. 2015). By definition, a pangenome represents all of the genetic potentials of a species and is typically determined by homology among sets of genes belonging to multiple strains of the species in all environments the species is found. Here, we extend the pangenome concept to incorporate metagenome-derived genes and genomes. It is a natural extension as MAGs and metagenomic contigs have been used to generate species-specific gene catalog and that for all species present in a given environment (Ma et al. 2019). We introduce the term, meta-pangenome that refers to the union of genes of a species found in a habitat using both culture-independent sequencing (metagenome) and culture-based sequencing (genome) methods. In computational terms, the meta-pangenome is the entire sequence space of a species in an environment. Thus, within a sample, a metagenomic species represents known combinations of strains of a species. In this chapter, we choose to discuss the meta-pangenome in the context of a species, while the meta-pangenome paradigm can be applied to genera or broader of taxonomic groups (Lefebure and Stanhope 2007) as well as other domains of life such as fungus (McCarthy and Fitzpatrick 2019). The term “pan” itself means “whole” or “everything”, and “meta” as a prefix could mean “with”, “among”, and “beyond”. Together the words “meta-pangenome” literally mean whole genomes of a species from among samples collected in a given environment.

Similar to the pangenome concept, a meta-pangenome is bound to a specific species. In order to define the meta-pangenome for a species, say species A, we start from collecting all available genomes and constructing MAGs of species A from metagenomes (illustrated in Fig. 1). We then perform gene calling for these MAGs contigs after quality assessment, followed by similarity search to generate homologous gene clusters as in conventional pangenome analyses. The final step is to perform meta-pangenome size interpolation and extrapolation for species A. This procedure can then be repeated for each of the species present in a particular environment to define their meta-pangenome. Alternatively, the genetic contents characterized in all metagenomes and genomes of a habitat can be collectively pooled to generate homologous gene clusters. Taxonomic assignment of the resulting gene clusters can then be used to produce meta-pangenomes for each of the species present in the habitat.

We can then apply the concepts of core, accessory, and unique genes to the meta-pangenome framework. A species meta-pangenome core genes are those consistently present in all or almost all metagenomes in a habitat such as wastewater or the GI tract, and meta-pangenome-specific genes are only observed a single sample of the habitat. The variable or accessory meta-pangenome includes those genes only present in a subset of populations. As a metagenome can be considered a snapshot of the microbial community genetic potential at the time of collection, the core meta-pangenome can be referred as the set of genes being repeatedly observed after

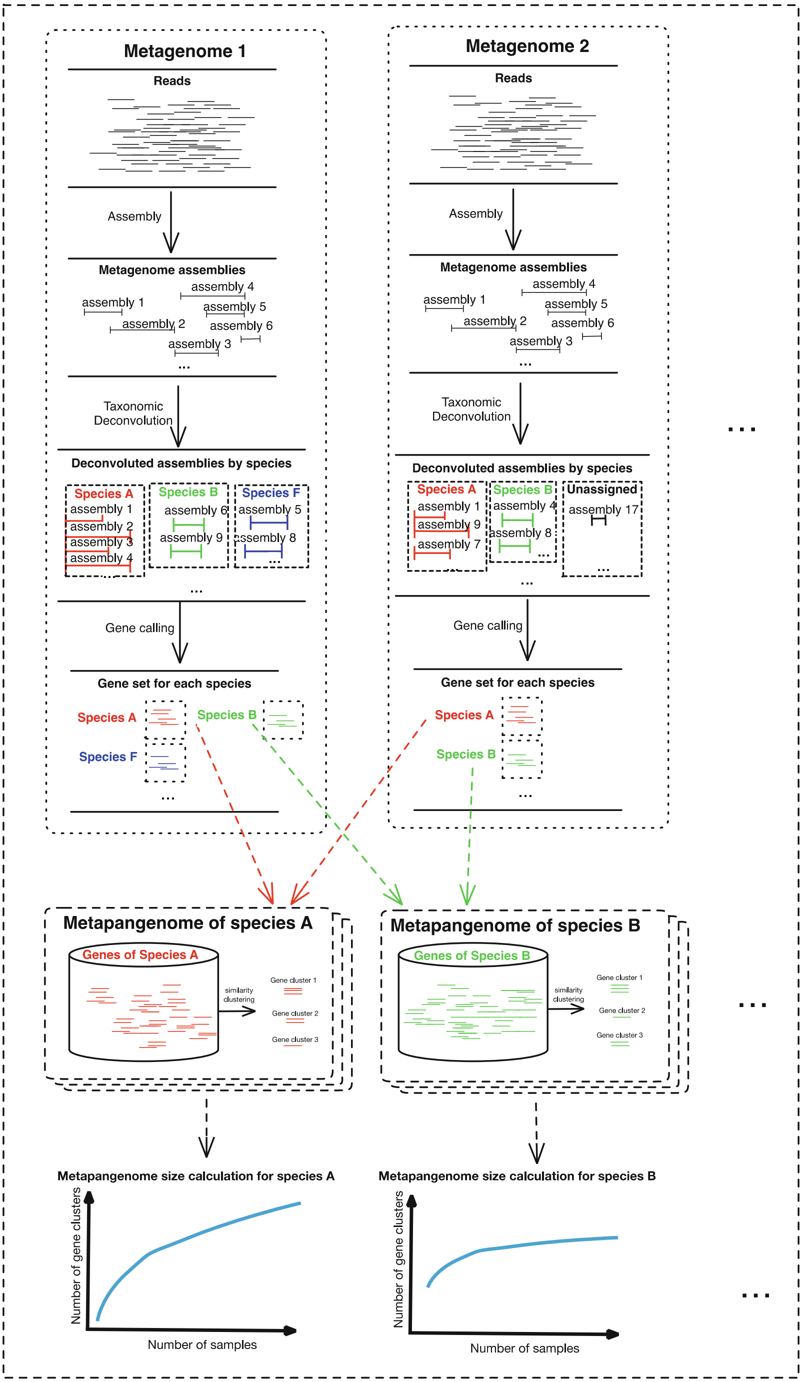
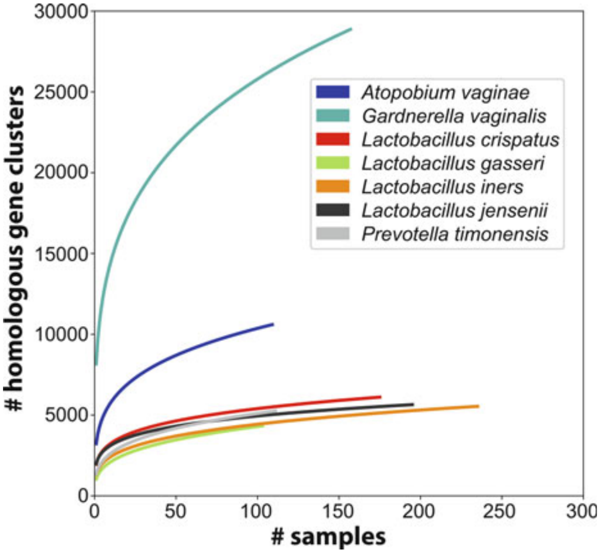


Fig. 1 Illustration for a workflow to generate a meta-pangenome for a species. The steps could be modified. For example, the step of gene calling could be after the step of the pooling all

Fig. 2 Species-specific metagenome accumulation curves for the number of homologous gene clusters. Figure reproduced from Ma et al. (2019)



multiple sampling events. A closed meta-pangenome would thus refer to the case where no or very few new genes of the species are added with each additional metagenome sequenced. Conversely, a species open meta-pangenome would refer to the case where a substantive number of new genes for that species are discovered with each additional metagenome sequenced. The core meta-pangenome for a species could be quite small, or even nonexistent, if the abiotic and biotic constraints on its colonization of the environment are loose or large if these constraints are strict.

Similar to the original pangenome ecological significance (Tettelin et al. 2005), population size and niche versatility are likely to drive the size of a meta-pangenome. For example, the meta-pangenome of *Gardnerella vaginalis*, a highly prevalent bacterial colonizer of human vagina, is a collection of all the genes assigned to that species derived from all available vaginal metagenomes and genomes. Despite hundreds of metagenomes available containing *G. vaginalis,* this important species shows an open meta-pangenome (Fig. 2). On the other hand, *Lactobacillus gasseri,* another important and beneficial vaginal bacterial species demonstrates an essentially closed meta-pangenome such that new metagenome sequences add relatively few genes. An in-depth understanding of the genetic diversity of constituent community members and its relation to community dysbiosis will afford the development of novel strategies to evaluate and optimize prevention, diagnostics, and treatment for adverse health conditions.

Fig. 1 (continued) deconvoluted assemblies for a species. Alternatively, the genetic contents characterized in all metagenomes and genomes of a habitat can be collectively pooled to generate homologous gene clusters. Taxonomic assignment of the resulting gene clusters can then be used to produce meta-pangenomes for each of the species present in the habitat

5 A Conceptual Framework for Microbial Comparative Genomics: Meta-Pangenome, Metagenomic Subspecies, and Pan-Metagenome

Meta-pangenome forms a practical framework that provides unprecedented insights into the genetic and functional basis underlying ecological fitness of microbial population in an environmental niche. The variable or accessory meta-pangenome of a species are the genes only present in a subset but not all of samples, which has led to the new concept of "metagenomic subspecies" (Ma et al. 2019). In essence, a metagenomic subspecies represents a slice of a species' meta-pangenome that is commonly identified in metagenomic samplings of a habitat. This slice contains the genetic contents of a combination of strains that tend to co-occur. In theory, this co-occurrence could be driven by interactions among the strains and/or their tendency to co-colonize, termed dispersal limitations (Telford et al. 2006). Specific mechanisms that can lead to the co-existence of multiple strains in a population include frequency-dependent selection (Svensson and Connallon 2019), cross-feeding (Livingston et al. 2012; Hunt and Bonsall 2009), spatial structure (France and Forney 2019), resource partitioning (Rosenzweig et al. 1994), and interference competition (Kerr et al. 2002), among others. That said, the metagenomic subspecies concept is equivalent to a species genetic "ecotype" for an environment. Several metagenomic subspecies can exist in a given environment but cannot co-occur within a sample. The metagenomic subspecies can be determined in silico by hierarchical clustering over the data matrix such as gene prevalence or gene abundance profiles. Further development of relevant pattern recognition tools (supervised or unsupervised) as well as the approximation of the population size (number of strains) are important ongoing research developments that will contribute to this field.

The concepts of meta-pangenome and metagenomic subspecies have great value to investigate intraspecies diversity within a community and the genetic foundation underlining the functions, resilience, resistance or fitness, among others, of microbial communities. We term the entire collection of all species' meta-pangenomes that exist in a specific environment the "pan-metagenome," which is essentially the "habitome" that encompasses the genetic landscape of a habitat. For instance, the pan-metagenome of the human gastrointestinal (GI) tract is the collection of all genes of all species found in the human GI tract (Qin et al. 2010; Li et al. 2014), and the pan-metagenome of the human oral communities encompasses the total genetic content of all species in the human oral environment (Tierney et al. 2019). The concept of pan-metagenome is represented by extensive gene cataloging, such as those constructed for the pig (Xiao et al. 2016) or the mouse GI tract (Xiao et al. 2015). A pan-metagenome of a specific habitat, when used as a catalog of the genetic contents, has provided a comprehensive reference framework for the study of microbial communities and their interaction with the environment.

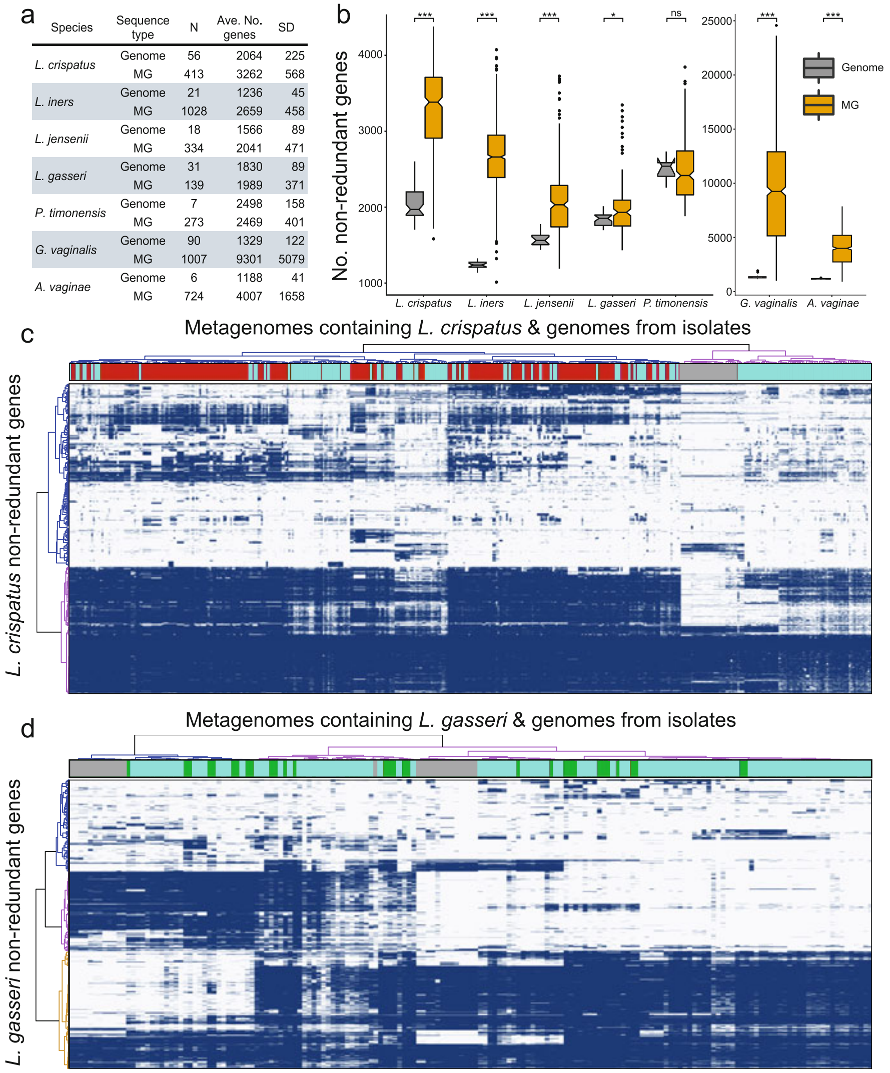
We have recently constructed a pan-metagenome for the human vaginal tract named VIRGO (the human vaginal nonredundant gene catalog) using an array of urogenital bacterial isolate genomes and vaginal metagenomes (Ma et al. 2019).

VIRGO has been shown to be comprehensive and to provide an unbiased representation of the genetic diversity of each species found in the vaginal microbiome. In building VIRGO, we found that the vast majority of the genetic diversity was contributed by MAGs derived from the metagenomic datasets. In fact, the metagenomic data used to build VIRGO comprise a much larger genetic diversity (high number of nonredundant genes) than that of all combined single isolate genome sequences (Fig. 3a, b). This result indicates the importance of extending the pangenome concept beyond isolate genome sequences.

VIRGO has afforded a different view of the vaginal microbiome, where each population is composed of complex mixtures of multiple strains, highlighting the large amount of intraspecies diversity present in these communities. We found that, in general, the majority of a species' genes are meta-pangenomic accessory genes. For example, for *Lactobacillus crispatus*, the number of meta-pangenomic accessory genes is twice as many as the number of meta-pangenomic core genes (Fig. 3c). *G. vaginalis* demonstrated particularly high intraspecies diversity, for which the core meta-pangenome does not even exist and the majority of the genes are accessory or sample specific, suggesting that the species should be split into multiple different species within the genus *Gardnerella.* We further observed three distinct metagenomic subspecies of *L. gasseri,* among which there were two distinct types and the third being a combination of the two (Fig. 3d). This suggests that there is environmental specialized co-colonization of *L. gasseri* strains in the vaginal environment. Future studies are needed to reveal the linkage between specific metagenomic subspecies and pathophysiological conditions.

6 Conclusion Remarks

The field of comparative genomics has bloomed from that initial genome comparison two decades ago. Thanks to advancements in cultivation-independent whole community sequencing technology and the increased availability of metagenome-assembled genomes, we have obtained unprecedented insights into the incredible amount of diversity present within microbial populations. Intraspecies diversity exceeds that found in our current reference genome databases. The pangenome paradigm expanded to metagenome-assembled genomes and metagenomic contigs comprehensively profile microbial genetic diversity in a specific habitat. However, the incorporation of metagenome-derived genomes has to be performed carefully with stringent quality assessment to avoid spurious inflation of gene content. The meta-pangenome concept unites pangenomics and metagenomics to obtain a more compete and ecologically meaningful view of different ecosystems. Meta-pangenomes and pan-metagenomes represent a critical step in the development of a systems-level understanding of microbial ecosystems.



Species	Sequence type	N	Ave. No. genes	SD
L. crispatus	Genome	56	2064	225
	MG	413	3262	568
L. iners	Genome	21	1236	45
	MG	1028	2659	458
L. jensenii	Genome	18	1566	89
	MG	334	2041	471
L. gasseri	Genome	31	1830	89
	MG	139	1989	371
P. timonensis	Genome	7	2498	158
	MG	273	2469	401
G. vaginalis	Genome	90	1329	122
	MG	1007	9301	5079
A. vaginae	Genome	6	1188	41
	MG	724	4007	1658

Fig. 3 Intraspecies diversity revealed using VIRGO (human vaginal nonredundant gene catalog) of seven vaginal species including *L. crispatus, L. iners, L. jensenii, L. gasseri*, and *G. vaginalis*, *A. vaginae* and *P. timonensis*. (**a**) Summary of the number (N) of isolate genomes and metagenome (MG) samples with more than 80% of their average genome's number of coding genes for a species, based on a dataset of 1507 *in-house* vaginal metagenomes characterized using VIRGO. (**b**) Boxplot of number nonredundant genes in isolate genomes versus vaginal metagenomes. (**c**) Heatmap of presence/absence of *L. crispatus* nonredundant gene profiles for 56 available isolate genomes and 413 VIRGO-characterized metagenomes that contained either high (red) or low (blue) relative abundance of the species. Hierarchical clustering of the profiles was performed using ward linkage based on their Jaccard similarity coefficient. ∗number of isolate genomes and metagenome samples. †MG: Metagenomes $*p < 0.05$, $***p < 0.001$ after correction for multiple comparisons. Figure reproduced from Ma et al. (2019)

Acknowledgment The author acknowledges The Gerber Foundation 2018 award.

Competing Interest Statement The authors declare no competing financial and nonfinancial interests.

References

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 31:533–538

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. Microbiol Rev 59:143–169

Bakken LR (1985) Separation and purification of bacteria from soil. Appl Environ Microbiol 49:1482–1487

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. Methods 58:268–276

Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J (2012) Ray meta: scalable de novo metagenome assembly and profiling. Genome Biol 13:R122

Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, Ngan CY, Cheng JF, Tringe SG, Woyke T (2015) Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics 16:856

Breitwieser FP, Lu J, Salzberg SL (2017) A review of methods and databases for metagenomic classification and assembly. Brief Bioinform 20(4):1125–1136

Costea PI, Munch R, Coelho LP, Paoli L, Sunagawa S, Bork P (2017) metaSNV: a tool for metagenomic strain level analysis. PLoS One 12:e0182392

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA (2005) Diversity of the human intestinal microbial flora. Science 308:1635–1638

Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319

Fitzsimons MS, Novotny M, Lo CC, Dichosa AE, Yee-Greenbaum JL, Snook JP, Gu W, Chertkov O, Davenport KW, McMurry K et al (2013) Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. Genome Res 23:878–888

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512

France MT, Forney LJ (2019) The relationship between spatial structure and the maintenance of diversity in microbial populations. Am Nat 193:503–513

France MT, Mendes-Soares H, Forney LJ (2016) Genomic comparisons of lactobacillus crispatus and lactobacillus iners reveal potential ecological drivers of community composition in the vagina. Appl Environ Microbiol 82:7063–7073

Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VGH, McHardy AC, Nederbragt AJ, Pope PB (2016) Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep 6:25373

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM et al (1995) The minimal gene complement of mycoplasma genitalium. Science 270:397–403

Gilbert JA, Dupont CL (2011) Microbial metagenomics: beyond the genome. Annu Rev Mar Sci 3:347–371

Greenblum S, Carr R, Borenstein E (2015) Extensive strain-level copy-number variation across human gut microbiome species. Cell 160(4):583–594

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68:669–685

Hardison RC (2003) Comparative genomics. PLoS Biol 1:E58

Howe A, Pell J, Canino-Koning R, Mackelprang R, Tringe S, Jansson J, Tiedje JM, Brown CT (2012) Illumina sequencing artifacts revealed by connectivity analysis of metagenomic datasets

Hunt JJ, Bonsall MB (2009) The effects of colonization, extinction and competition on co-existence in metacommunities. J Anim Ecol 78:866–879

Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. Genome Res 21:1552–1560

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335:587–590

Kerr B, Riley MA, Feldman MW, Bohannan BJ (2002) Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. Nature 418:171–174

Koren S, Phillippy AM (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 23:110–120

Kraal L, Abubucker S, Kota K, Fischbach MA, Mitreva M (2014) The prevalence of species and strains in the human microbiome: a resource for experimental efforts. PLoS One 9:e97279

Lefebure T, Stanhope MJ (2007) Evolution of the core and pan-genome of streptococcus: positive selection, recombination, and genome composition. Genome Biol 8:R71

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272

Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T et al (2014) An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol 32:834–841

Li D, Liu CM, Luo R, Sadakane K, Lam TW (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676

Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, Yamashita H, Lam TW (2016) MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102:3–11

Livingston G, Matias M, Calcagno V, Barbera C, Combe M, Leibold MA, Mouquet N (2012) Competition-colonization dynamics in experimental bacterial metacommunities. Nat Commun 3:1234

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1(1):18

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D (2015) ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol 33:1045–1052

Ma B, France M, Crabtree J, Holm J, Humphrys M, Brotman R, Ravel J (2019) VIRGO, a comprehensive non-redundant gene catalog, reveals extensive within community intraspecies diversity in the human vagina. bioRxiv

Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, Jansson JK (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature 480:368–371

McCarthy CGP, Fitzpatrick DA (2019) Pan-genome analyses of model fungal species. Microb Genom 5:e000243

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15:589–594

Meyer F, Hofmann P, Belmann P, Garrido-Oter R, Fritz A, Sczyrba A, McHardy AC (2018) AMBER: assessment of Metagenome BinnERs. Gigascience 7

Miller W, Makova KD, Nekrutenko A, Hardison RC (2004) Comparative genomics. Annu Rev Genomics Hum Genet 5:15–56

Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 40:e155

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017) metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834

Oh J, Byrd AL, Deming C, Conlan S, Program NCS, Kong HH, Segre JA (2014) Biogeography and individuality shape function in the human skin metagenome. Nature 514:59–64

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015) Check M: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055

Pell J, Hintze A, Canino-Koning R, Howe A, Tiedje JM, Brown CT (2012) Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proc Natl Acad Sci U S A 109:13272–13277

Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428

Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn LJ, Knetsch CW, Figueiredo C (2019) Impact of host DNA and sequencing depth on the taxonomic resolution of whole Metagenome sequencing for microbiome analysis. Front Microbiol 10:1277

Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A 98:9748–9753

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65

Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, Eren AM (2017) DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol 18:181

Rosenzweig RF, Sharp RR, Treves DS, Adams J (1994) Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*. Genetics 137:903–917

Sangwan N, Xia F, Gilbert JA (2016) Recovering complete and draft population genomes from metagenome datasets. Microbiome 4:8

Shaiber A, Eren AM (2019) Composite metagenome-assembled genomes reduce the quality of public genome repositories. MBio 10(3):e00725–e00719

Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF (2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol 3:836–843

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proc Natl Acad Sci U S A 103:12115–12120

Stewart RD, Auffret MD, Snelling TJ, Roehe R, Watson M (2019) MAGpy: a reproducible pipeline for the downstream analysis of metagenome-assembled genomes (MAGs). Bioinformatics 35:2150–2152

Svensson EI, Connallon T (2019) How frequency-dependent selection affects population fitness, maladaptation and evolutionary rescue. Evol Appl 12:1243–1258

Telford RJ, Vandvik V, Birks HJ (2006) Dispersal limitations matter for microbial morphospecies. Science 312:1015

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS 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

Tierney BT, Yang Z, Luber JM, Beaudin M, Wibowo MC, Baek C, Mehlenbacher E, Patel CJ, Kostic AD (2019) The landscape of genetic content in the gut and Oral human microbiome. Cell Host Microbe 26:283–295. e288

Touchman J (2010) Comparative genomics. Nat Educ Knowl 3:13

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N (2017) Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res 27:626–638

Tsai YC, Conlan S, Deming C, Program NCS, Segre JA, Kong HH, Korlach J, Oh J (2016) Resolving the complexity of human skin metagenomes using single-molecule sequencing. MBio 7:e01948–e01915

Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154

Xia X (2013) Comparative genomics. In Briefs in Genetics. Springer, Heidelberg

Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, Li X, Long H, Zhang J, Zhang D et al (2015) A catalog of the mouse gut metagenome. Nat Biotechnol 33:1103–1108

Xiao L, Estelle J, Kiilerich P, Ramayo-Caldas Y, Xia Z, Feng Q, Liang S, Pedersen AO, Kjeldsen NJ, Liu C et al (2016) A reference gene catalogue of the pig gut microbiome. Nat Microbiol 1:16161

Zaheer R, Noyes N, Ortega Polo R, Cook SR, Marinier E, Van Domselaar G, Belk KE, Morley PS, McAllister TA (2018) Impact of sequencing depth on the characterization of the microbiome and resistome. Sci Rep 8:5890

Pangenome Flux Balance Analysis Toward Panphenomes

Charles J. Norsigian, Xin Fang, Bernhard O. Palsson, and Jonathan M. Monk

Abstract Studies of the pangenome have been empowered by an exponentially increasing amount of strain-specific genome sequencing data. With this data deluge comes a need for new tools to contextualize, analyze, and interpret such a vast amount of information. Network reconstructions, genome-scale metabolic models (GEMs), and the corresponding computational analysis frameworks such as flux balance analysis (FBA) have been proven useful toward this end. Network reconstructions can be used to interpret genomic variation not just from a single strain but for an entire species. By applying these approaches at the pangenome scale, it becomes possible to systematically evaluate phenotypic properties for an entire species thus enabling the study of a panphenome directly from a pangenome. Applying insights gained from analysis of the panphenome has diverse implications with applications ranging from human health to metabolic engineering. The future of pangenomics will include panphenomic analyses, thus supplementing traditional pangenomic analyses and helping to address the Big-data-to-knowledge grand challenge of analyzing thousands of genomic sequences.

Keywords Flux balance analysis · Genome-scale modeling · Panphenome · Multi-strain · Comparative systems biology

1 Introduction

Studying differences between strains of a species using the construct of a pangenome revolutionized the field of comparative genomics for bacteria (Tettelin et al. 2005; Medini et al. 2005). This framework allowed scientists to overcome issues related to species with high genomic variability and lack of a reference genome. The pangenome alone cannot be used to quantify the phenotypic effects

C. J. Norsigian · X. Fang · B. O. Palsson · J. M. Monk (✉)
Department of Bioengineering, University of California San Diego, La Jolla, CA, USA
e-mail: jmonk@ucsd.edu

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_10

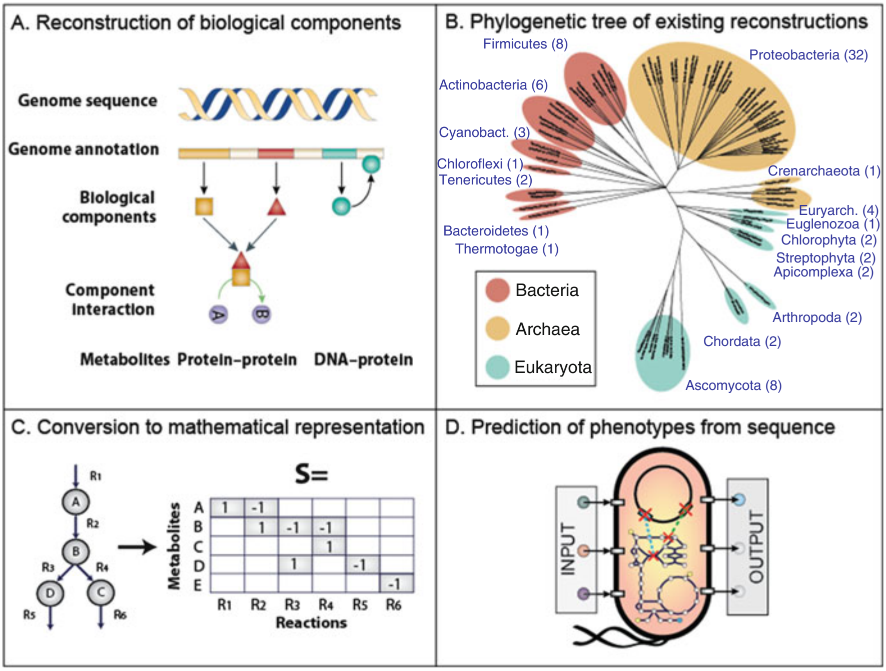
of genetic variability. Over the past decade, network reconstructions have become an indispensable tool in molecular systems biology because of their ability to provide a mechanistic link between experimental studies and computational analyses (Bordbar et al. 2014). Thus, genome-scale network reconstructions provide an avenue for extending the power of the pangenome toward evaluating the phenotypic capabilities of a species or the panphenome. High-quality reconstructions can be expanded through bioinformatic techniques to map information from a reference strain to additional strains of the target organism. This chapter describes how reconstructions and genome-scale models have been applied to study the pangenome by predicting all possible phenotypes for strains in a species. Using these tools, large-scale genomic data sets combined with experimental phenotypes can now be integrated and queried to systematically probe the diversity of strains within a species. Genome-scale metabolic network reconstructions can delineate conserved and unique metabolic capabilities across the strains of a species. These differences and designations can be used to define the metabolic potential of a species often informative of lifestyle diversity. In this chapter, we detail the following elements toward true panphenomic analysis: (1) The foundation of reconstructions and flux balance analysis; (2) The extension of these tools using a "multi-strain" approach to calculate metabolic panphenomes for several bacterial species; and (3) A future perspective on the multi-strain approach: moving beyond metabolism for a full calculation of the panphenome.

2 Network Reconstructions and Flux Balance Analysis

The growing collections of sequences that have been used to study pangenomes are laden with valuable information, however, strings of nucleotide bases alone do not make this information easily accessible or immediately apparent. Thus, there is a critical need for tools that can be used to interrogate this massive amount of data to generate new knowledge. Genome-scale network reconstructions in concert with flux balance analysis (FBA) provide such a tool. This section describes the process of reconstruction as well as mathematical approaches that can be used to query and compute with reconstruction, in particular, FBA.

2.1 Network Reconstructions Structure Biological Knowledge

Genome-scale reconstructions are organism-specific knowledge bases. They are built systematically using a quality-controlled bottom-up workflow that incorporates genome annotation, omics data sets, and legacy knowledge. The literature detailing the construction and analysis of network reconstructions is extensive (O'Brien et al. 2015; Thiele and Palsson 2010; Herrgård et al. 2008). In brief, these tools organize knowledge by linking genes, gene products, and cellular components (Fig. 1a).



	R1	R2	R3	R4	R5	R6
A	1	-1				
B		1	-1	-1		
C				1		
D			1		-1	
E						-1

Fig. 1 (**a**) Reconstructions consist of layered information connecting annotated genes on the genome sequence to their encoded biological products (e.g., RNA, protein) and how those components interact with other biological components (e.g., protein metabolite, in the case of a metabolic reaction/transformation. Figure reprinted from Reed et al. (2006). (**b**) Genome-scale models exist for species across the tree of life that are being made for new species and constantly improving. Reprint from Monk et al. (2014). (**c**) Reconstructions can be converted to a mathematical format by account for use of biological components (e.g., consumption/production). This allows for molecular accounting and enforcement of constraints. (**d**) Enforcement of constraints (e.g., media updates) and applying an objective (e.g., production of biomass, e.g., growth) allows for simulation of biological phenotypes from the genotype. Panel c and d reproduced from O'Brien et al. (2015). Reprint from O'Brien et al.)

Reconstructions can be made for several cellular processes including transcriptional regulation (Gianchandani et al. 2006, 2009), expression (Thiele et al. 2009) and metabolism (Feist et al. 2009). The reconstruction approach is iterative and thus all reconstructions are continually improving as new knowledge is generated. Thus, reconstructions serve as a valuable resource to integrate and reconcile biochemical data allowing researchers to collaborate, test, and readily share new hypotheses about functions in a target organism (Monk et al. 2014).

Reconstructions of cellular metabolism have been the most developed and extensively used type thus far (Bordbar et al. 2014). Metabolic network reconstructions are composed of all known metabolic genes, their encoded proteins and catalyzed reactions. This information is synthesized by aggregating organism-specific databases, high-throughput data, and primary literature (Thiele and Palsson 2010).

Advancements have allowed for partial automation of this process (Henry et al. 2010; Agren et al. 2013). Reactions are organized into pathways, pathways into subsystems, and ultimately into genome-scale networks; thus, representing biological processes at multiple scales. The resulting network reconstruction is a unification of the information available for an organism with a genetic basis. Today, there exist collections of genome-scale reconstructions for a number of target organisms across the tree of life (Oberhardt et al. 2011; Monk et al. 2014) (Fig. 1b). For example, as of 2018, there are 178 available, curated reconstructions spanning the tree of life (http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms). While this coverage is impressive, several other phyla remain devoid of any reconstruction initiative. To fully extend the study of panphenomes to all sequenced organisms, new reconstruction efforts must be initiated (Monk et al. 2014).

2.2 *Flux Balance Analysis Enables Computation of Phenotype from Genotype*

Reconstructions alone are static, and unable to be used for predictions. A major value of the metabolic reconstructions emerges when they are converted into a mathematical format, enabling computational interrogation using a variety of methods (Orth et al. 2010; Lewis et al. 2012). This conversion translates the biochemical reactions of a reconstructed network via tabulation of reaction stoichiometry into a chemically accurate mathematical format that becomes the basis for a genome-scale model (GEM) (Fig. 1c). The flow of metabolites through the network is constrained by these stoichiometries represented as balances or inequalities for bounds (Reed 2012). Further constraints can be added to a network such as thermodynamic reversibility constraints and limitations to nutrient uptake or by-product secretion. Computationally predicted network states consistent with imposed constraints are potential physiological states of the target organism within a defined condition.

Flux balance analysis (FBA) can be applied to these models for prediction of an organism's phenotype. This mathematical approach for analyzing the flow of metabolites through a metabolic network is the original constraints-based method (Orth et al. 2010). This approach relies on an assumption of steady-state growth and mass balance. FBA uses the stated objective (for example, biomass production, e.g., growth) to find the solution(s) using linear programming that optimize an objective function (O'Brien et al. 2015). In a defined environment (defined inputs), GEMs can be used to compute network outputs (Fig. 1d) FBA allows for computational tracing of balanced reaction states beginning with defined inputs to produce output metabolites. Biomass synthesis is computed using FBA by computing the balanced reactions states that produce all the required metabolites for growth simultaneously. Additionally, the model accounts for the energetic, redox, and chemical balances that must also be maintained (O'Brien et al. 2015).

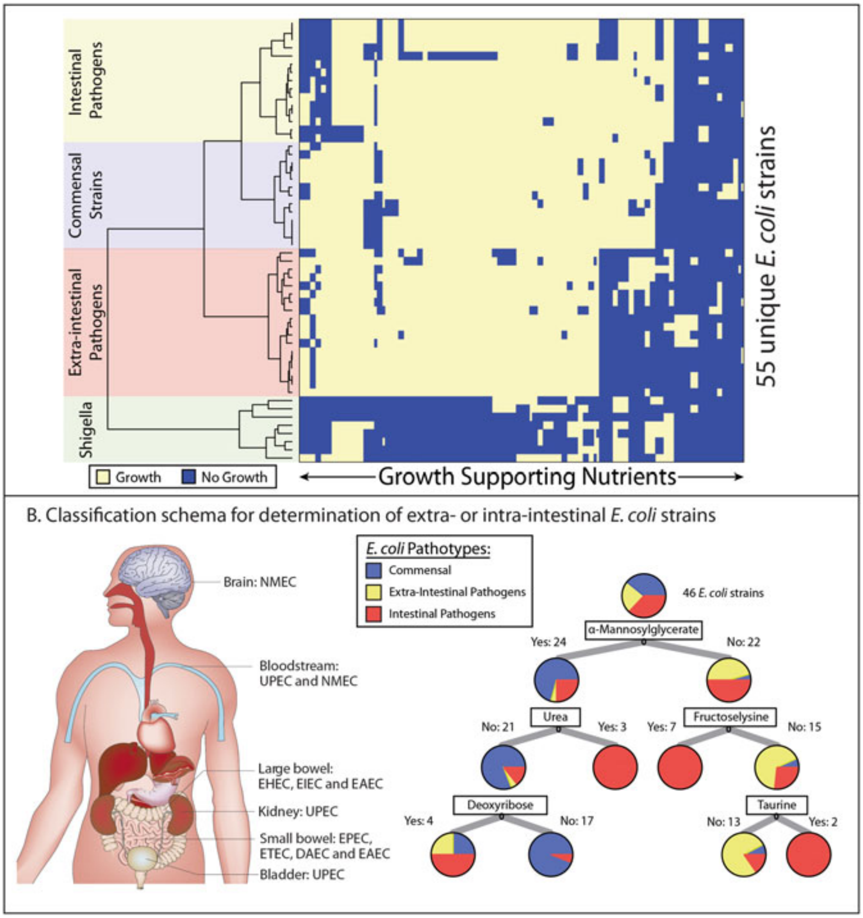
Using this technique, a variety of phenotypes such as the effect of gene knockouts, metabolite secretion, and growth capabilities on different substrates can be predicted rapidly and compared to experimental results to verify their accuracy (Monk and Palsson 2014). Some of the best models have accuracies >90% in agreement with experimental data (Monk et al. 2017; Brunk et al. 2018). In this way, GEMs provide a way to bridge the genotype to phenotype gap by providing a robust platform for analyzing the integrated mechanisms of gene products to produce unique phenotypic states. The utility of a highly curated GEM and the corresponding computational analyses is increased by the format's scalability. Through this methodology, phenotypes for the plethora of sequenced strains within a species become readily calculable. In the next section, we will highlight how high-quality reconstructions for a single strain can be extrapolated onto several strains of the same species to study the phenotypic potential of the pangenome and to gain insight into strain-specific metabolic capabilities.

3 The Multi-Strain Approach: Extending Genome-Scale Models to Robustly Explore the Pangenome Phenotypic Space

Once a high-quality reconstruction and genome-scale model exist, its contents (e.g., genes, metabolites, and reactions) can be mapped onto other, closely related strains in a species. Following this multi-strain approach, tools from comparative genomics (Monk and Bosi 2018) can be integrated with genome-scale modeling to identify genetic determinants underlying variability of phenotypes. Such a task is crucial to understand the evolutionary trajectories of a bacterial species. Strain-specific metabolic diversity has been illuminated through the use of genome-scale metabolic models. Prediction of unique metabolic capabilities and auxotrophies can be used to study species lifestyle diversity. This approach is scalable to the pangenome level and in turn enables panphenome analysis, thus empowering species-wide comparative systems biology. This multi-strain approach has been applied to several species in a variety of studies and we provide a brief overview of the key insights here.

3.1 Genesis of the Multi-Strain Approach: Studying Escherichia coli

The first instance of the multi-strain approach as described here was executed by Monk et al. where the authors leveraged a curated genome-scale model of *E. coli* K-12 MG1655 that has been continually updated over 15 years to construct genome-scale models of 55 other fully sequenced *E. coli* strains (Monk et al. 2013). Using FBA on all 55 of these models, the authors were able to extensively investigate the



B. Classification schema for determination of extra- or intra-intestinal *E. coli* strains

Fig. 2 (**a**) Genome-scale models can be used to predict growth capabilities in different environments and nutritional niches. This figure represents growth predictions for 55 different strain-specific models of *E. coli* and *Shigella* on over 300 different carbon, nitrogen, phosphorus and sulfur sources. Strains, for the most part, clustered according to their isolated niches (e.g., extra versus intestinal). Reproduced from Monk et al. (2013). (**b**) Using these growth predictions allows for the classification of strains and their potential isolation site (e.g., bladder versus intestine). Decision trees could reliably separate ExPEC from InPEC strains. Left panel reproduced from Croxen and Finlay (2010). Right panel reproduced from Monk et al. (2013)

predicted metabolic capabilities of all the strains (Fig. 2a). The authors delineated strain-specific auxotrophies and substrate preferences among the set of strains. It is important to note that these predictions and insights were gained from sequence alone. Further, this study demonstrated the possibility of applying this approach to understand cases of patho-adaptation to a given environment and evaluate a given strain's infectious niche.

Further work scaled up the effort to include 1200 strains of *E. coli* and demonstrated a large amount of variability within the species both in gene content and consequent variability of gene products (Monk et al. 2017). It also utilized the differences across the 1200 strains to construct a robust classification tree for determination between extra-intestinal and intra-intestinal pathogens using predicted metabolic phenotypes (Fig. 2b). This type of classification schema opens the door to investigating how strain-specific traits impact the microbiome. An in-depth example of such analyses came in a study by Fang et al. into the metabolic capabilities of inflammatory bowel disease (IBD)-associated *E. coli* strains in the B2 clade (Fang et al. 2018). The authors found these strains have advantages in catabolizing sugars derived from mucus glycans. The interesting and novel outcomes of these *E. coli* studies clearly demonstrated the value of the approach, and the natural next step was to apply the methodology to other species.

3.2 Expanding the Reach of Multi-Strain Approach Across the Phylogenetic Tree

Numerous studies followed the first *E. coli* studies that focused on various organisms. Fouts et al. applied the multi-strain approach, broadened to examine various species of *Leptospira* known to have ranging levels of pathogenicity (Fouts et al. 2016). They demonstrated that the ability to synthesize vitamin B12 is limited to pathogenic species of *Leptospira* and may give them a survival advantage in a human host where B12 is sequestered by the body. This valuable distinguishing metabolic capability was captured by being able to leverage the base reconstruction across multiple species in the genus.

In 2016, Bosi et al. applied the workflow to 64 strains of *Staphylococcus aureus*. Beyond reconstructing metabolic capabilities, the approach was extended to identify virulence factors in the set of 64 strains (Bosi et al. 2016). By using a combination of predicted metabolic capabilities linked to virulence factors, they were able to stratify the strains by host type. This study added an additional layer to the promise of the multi-strain approach by showing that metabolic capabilities could be analyzed in concert with other components of the pangenome, namely virulence factors (toxins, adhesins, etc.), and that this combination held predictive power about a strain's host. This study also included explicit calculation of the core- and pangenome content of *S. aureus*, a metric of genomic diversity among strains in a species.

The multi-strain approach has also been applied to other pathogens such as *Acinetobacter baumannii* and *Salmonella.* In a study by Norsigian et al., a highly curated base GEM was used to create models for 75 different *A. baumannii* strains (Norsigian et al. 2018). These strain-specific models demonstrated major differences in metabolism between strains indicating that a classification scheme may be possible from sequence alone. Seif et al. built strain-specific models for 450 *Salmonella* strains from various serovars to show that metabolic capabilities can be used to distinguish these serovars (Seif et al. 2018). This study indicates that the host-range may be limited by metabolic capabilities of different strains.

3.3 Extending the Multi-Strain Approach to Investigate Additional Biological Qualities

The multi-strain framework provides an inherently efficient means of interrogating the properties of many strains and a few studies have utilized this organizational efficiency to gain insight into properties outside of direct metabolic capabilities. For example, Choudhary et al. examined the agr type of 400 *S. aureus* strains to examine the structure of genes within the genome (Choudhary et al. 2018). The authors found that genomic virulence factor profiles are highly correlated with agr type. They also identified that divergence in histidine kinase protein confers signal specificity with clear differences in protein structural properties based on agr types. Another example of additional properties is the investigation of reactive oxygen species (ROS) tolerance. By leveraging the multi-strain approach in conjunction with 3D structures Mih et al. was able to simulate ROS production levels to demonstrate that antioxidant properties are exhibited in the structural proteome (Mih et al. 2018). A third example was conducted by Kavvas et al., who took a deeper level of resolution within the genome by looking at the unique alleles present within *Mycobacterium tuberculosis* genomes (Kavvas et al. 2018). Through machine learning techniques on the pangenome they were able to associate certain alleles potentially responsible for antimicrobial resistance. The results hint at metabolic rewiring at the allelic level required for adaptation to antibiotic resistance. The success of the multi-strain approach in all these various studies suggests that explicit calculation of the panphenome will provide novel insights.

4 Future Perspectives: Moving Beyond Metabolism: A Multi-Scale Approach to Calculating Full Panphenomes

This chapter details a computational approach (network reconstruction and FBA) to systematically calculate metabolic phenotypes for multiple strains in a species. Beyond calculation of metabolic phenotypes, new methods, both experimental and computational, offer exciting new avenues for research into the pangenome. These approaches can be applied at multiple different scales. At the lowest level, single nucleotide variants (SNV) can be compared across strains using sequence mapping toolkits like breseq and gatk (Deatherage and Barrick 2014; McKenna et al. 2010). These approaches can be scaled up from single base changes to full gene sequences to compare orthologous ORFs across genomes by comparing sequence-specific alleles across strains in a species (Fig. 3a). As described here, the presence/absence of given

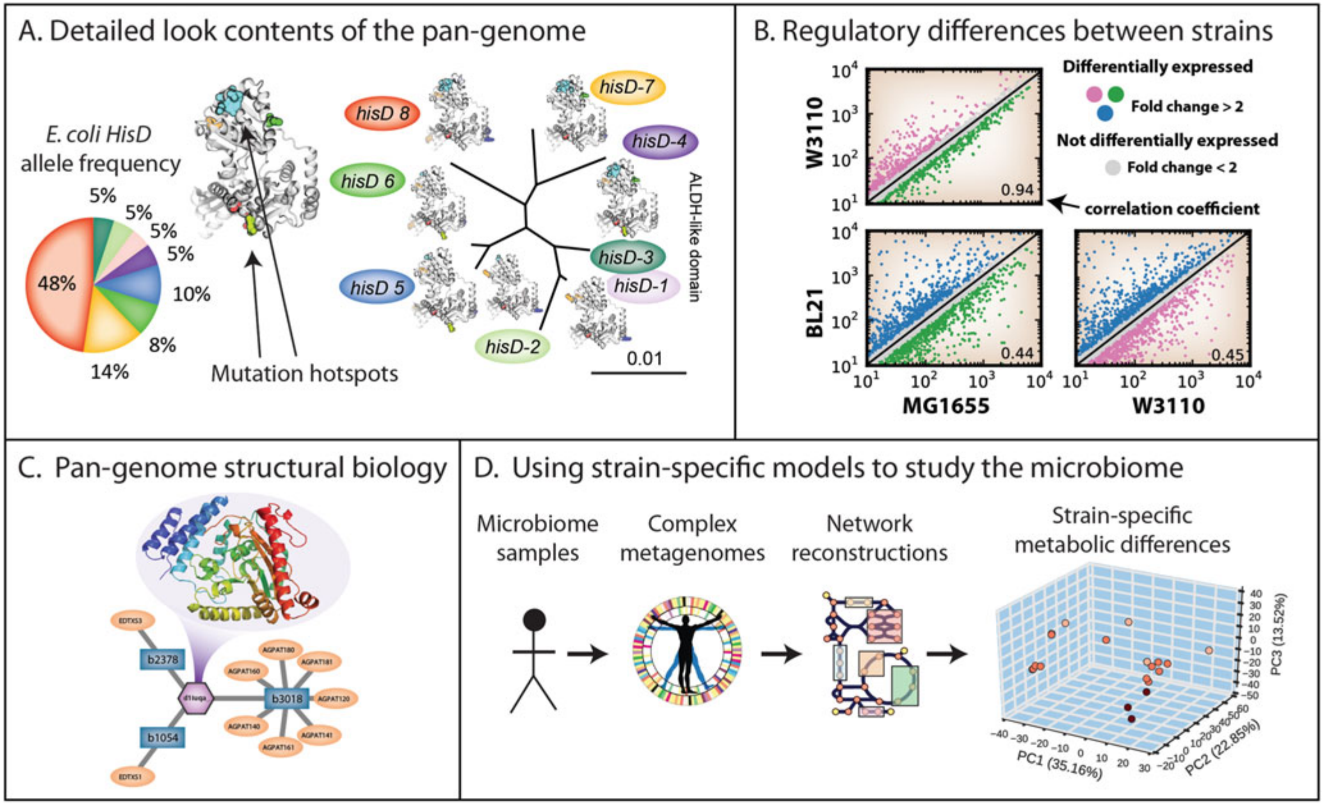


Fig. 3 (**a**) Detailed view of amino acid polymorphisms (allele frequency) for this D gene among 1200 diverse *E. coli* strains. Phylogenetic tree illustrating the relatedness between the unique alleles. (**b**) Comparison of orthologous gene expression between three different strains of *E. coli* (K-12 W3110, K-12 MG1655

enzyme-encoding metabolic genes can be used to build strain-specific metabolic reconstructions that compute metabolic panphenomes. While most of the applications described here are applied to pathogens with relevance to human health, it is important to note that the pangenome can also be studied for use in metabolic engineering applications. For example, the pangenome can be mined to search for enzymes of interest to industrial microbiology (Moscatello and Pfeifer 2018).

In the future, processes beyond metabolism will also be reconstructed allowing for true panphenome calculations. For example, reconstructions of protein expression mechanisms already exist (Thiele et al. 2009) and have been integrated with models of metabolism (ME models) (O'Brien et al. 2013). These models account for the transcription and translation processes and molecular constituents required to express enzymes catalyzing metabolic reactions in the metabolic network. It is further possible to use the ME model framework to reconstruct proteostatic mechanisms and investigate the structural integrity of the proteome (Chen et al. 2017). In the future, multiple ME models of strains in a species will further expand the scope of computation possible on contents of the pangenome.

Beyond metabolism and expression, regulatory networks are another aspect of the pangenome that differ between strains and have been reconstructed for individual strains (Gianchandani et al. 2006, 2009). Understanding how certain strains regulate the same set of genes (core-genome), as well as diverse sets of genes, will further expand our understanding of the structure and function of the pangenome. A small-scale study of seven *E. coli* strains and their RNA-seq expression profiles in aerobic and anaerobic environments showed remarkably different expression levels even for shared genes of the core-genome (Monk et al. 2016) (Fig. 3b). Studying differentially expressed genes and the transcription factors known to regulate them may lead to the discovery of alternative regulatory strategies between strains of a species.

Just as sequence databases have grown tremendously in recent years, 3D crystal structures for the encoded genes have also grown dramatically (Brunk et al. 2016). The protein data bank (Berman et al. 2000) (PDB) is a repository of protein structures and these structures can now be integrated with genome-scale models (GEM-PRO) (Chang et al. 2013). Building multi-strain models with associated protein structures is another way to compare strains across a species. Using these tools, sequence diversity can be examined at the 3D level to see how mutations line up in 3D space, a level of analysis not possible at the sequence level. Furthermore, mutations in specific regions of the protein can be tabulated (Fig. 3c) and compared across strains (Mih et al. 2018).

Finally, a multi-strain approach should prove useful for studies of the microbiome. Multiple genome-scale models for species found in the microbiome

Fig. 3 (continued) and BL21). Overall the K-12 strains have a much higher correlation between their transcriptional profiles than did BL21. Reproduced from Monk et al. (2016) (**c**) Expanding analysis of sequence similarity by incorporating 3D structural information. The inclusion of structures mapped to sequences allows the visualization of how differences in sequences manifest in 3D space. (d) Expanding study of strains to the microbiome using metagenomics and strain-level resolution. Panels a, c, and d reproduced from Monk et al. (2017)

already exist (Magnúsdóttir et al. 2017), and GEM studies were proven effective in studying the impact of diet (Shoaie et al. 2015) and interactions between microbes (Shoaie et al. 2013). Expanding the multi-strain approach to study diverse strains in these species may lead to a deeper level understanding of the gut microbiome composition. Indeed, strain-level metagenomics is coming (Scholz et al. 2016) and expanding the study of the pangenome to the microbiome will have fruitful applications in the near future (Fig. 3d).

In closing, we must list some caveats and risks to the multi-strain approach. First, all of these approaches require high-quality sequence data connected to high quality, QC/QA data generation. The success of reliable and maximally effective future panphenomics rests on ensuring this quality. There must be a continued effort to ensure that sequencing projects are of quality not only quantity. Additionally, an interesting question pertaining to the concept of closed pangenomes is, how will the law of diminishing returns be exhibited in these sequence deposits? Will a point be reached where additional sequences provide no novel information? Further, the vision of the panphenome and its implications to understanding how microbial pathogens impact human health will rely on both the availability of metadata and the deposition of strains. Metadata on these strains will only deepen the possible questions to be asked of both pangenomes and panphenomes. A centralized repository of strains will also greatly expedite the experimental verification needed for such large computational predictions. The future of the panphenome is apparent and with it further explanations at the center of biological causality.

5 Conclusions

Significant advancements in DNA sequencing technology have led to an exponential increase in the number of sequenced strains. This creates a need for new ways to integrate and analyze this ever-increasing amount of sequence information. This need will only intensify as the number of sequenced strains within a species continues to grow exponentially. This chapter demonstrates how the pangenome is evolving from a theoretical concept to a queryable construct.

In this chapter, we describe how the foundational aspects of GEMs and FBA can be used to predict phenotypic states for multiple strains in a species. The multi-strain approach has proven useful in extending this utility in a number of studies providing evolutionary insights as well as practical applications. As the library of available sequences continues to grow, the possibility of scaling these techniques to the level of the pangenome is becoming a reality. The result, a species-wide panphenome, would create a deeper level of understanding than the collection of gene content within the pangenome alone.

The ability to systematically characterize an entire species' phenotypic capabilities will enhance the depth of pangenome analysis possible and pull valuable information inherent to genome sequences to the forefront (Fig. 4). The linkages and distinct features at the pangenome scale for a species offer obvious value for future knowledge generation, especially pertaining to human health and disease.

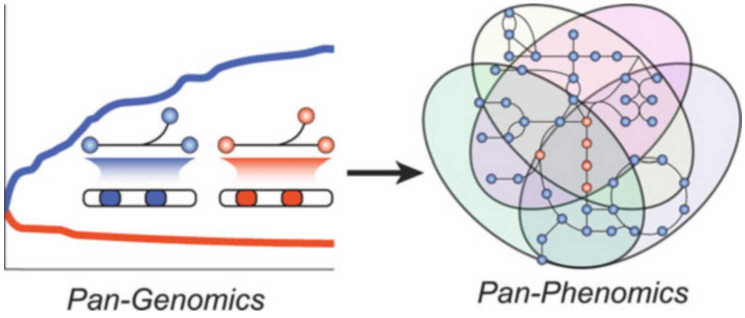


Fig. 4 The established assembly of the pangenome through the use of genome-scale reconstructions and corresponding computational analyses enables the calculation of panphenomes. The panphenome increases the depth of analysis possible by providing a framework in which to delineate strain-specific phenotypes. This stratification based on sequence similarity allows for the determination of which pieces of reconstructed networks are shared among various groups of strains in a species. This will continue to further inform the generation of evolutionary hypotheses

Further, the future potential applications outlined here such as inclusion of expression, regulation, and structures into these workflows will only further advance the scope of genome-scale science. Genome sequences are laden with critical information and the tools/workflows described in this chapter provide a means for extracting this information into actionable knowledge.

References

Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. PLoS Comput Biol 9(3):e1002980

Berman HM, Westbrook J, Feng Z (2000) The protein data bank. Nucleic Acids Res 28 (1):235–242. https://academic.oup.com/nar/article/doi/10.1093/nar/28.1.235/2384399/The-Protein-Data-Bank

Bordbar A, Monk JM, King ZA, Palsson BO (2014) Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet 15(2):107–120

Bosi E, Monk JM, Aziz RK, Fondi M, Nizet V, Palsson BØ (2016) Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. Proc Natl Acad Sci USA 113(26):E3801–E3809

Brunk E, Mih N, Monk J, Zhang Z, O'Brien EJ, Bliven SE, Chen K, Chang RL, Bourne PE, Palsson BO (2016) Systems biology of the structural proteome. BMC Syst Biol 10(March):26

Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, Mih N, Gatto F et al (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat Biotechnol 36 (3):272–281

Chang RL, Andrews K, Kim D, Li Z, Godzik A, Palsson BO (2013) Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli*. Science 340(6137):1220–1223. https://doi.org/10.1126/science.1234012. PMID: 23744946

Chen K et al (2017) Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. Proc Natl Acad Sci 114(43):11548–11553

Choudhary KS, Mih N, Monk J, Kavvas E, Yurkovich JT, Sakoulas G, Palsson BO (2018) The *Staphylococcus aureus* two-component system AgrAC displays four distinct genomic arrangements that delineate genomic virulence factor signatures. Front Microbiol 9:1082

Croxen MA, Finlay BB (2010) Molecular mechanisms of *Escherichia coli pathogenicity*. Nat Rev Microbiol 8(1):26–38

Deatherage DE, Barrick JE (2014) Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using Breseq. Methods Mol Biol 1151:165–188

Fang X, Monk JM, Mih N, Du B, Sastry AV, Kavvas E, Seif Y, Smarr L, Palsson BO (2018) *Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. BMC Syst Biol 12(1):66

Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7(2):129–143

Fouts DE, Matthias MA, Adhikarla H, Adler B, Amorim-Santos L, Berg DE, Bulach D et al (2016) What makes a bacterial species pathogenic? Comparative genomic analysis of the genus Leptospira. PLoS Negl Trop Dis 10(2):e0004403

Gianchandani EP, Papin JA, Price ND, Joyce AR, Palsson BO (2006) Matrix formalism to describe functional states of transcriptional regulatory systems. PLoS Comput Biol 2(8):e101

Gianchandani EP, Joyce AR, Palsson BØ, Papin JA (2009) Functional states of the genome-scale *Escherichia coli* transcriptional regulatory system. PLoS Comput Biol 5(6):e1000403

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28 (9):977–982

Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2008) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7:129–143. http://www.nature.com/nrmicro/journal/v7/n2/abs/nrmicro1949.html

Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, Heckmann D et al (2018) Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. Nat Commun 9(1):4306

Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10(4):291–305

Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K et al (2017) Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol 35(1):81–89

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15(6):589–594

Mih N, Brunk E, Chen K, Catoiu E, Sastry A, Kavvas E, Monk JM, Zhang Z, Palsson BO (2018) ssbio: a python framework for structural systems biology. Bioinformatics 34(12):2155–2157

Monk J, Bosi E (2018) Integration of comparative genomics with genome-scale metabolic modeling to investigate strain-specific phenotypical differences. In: Fondi M (ed) Metabolic network reconstruction and modeling: methods and protocols. Springer, New York, pp 151–175

Monk J, Palsson BO (2014) Genetics. Predicting microbial growth. Science 344(6191):1448–1449

Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, Feist AM, Palsson BØ (2013) Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. Proc Natl Acad Sci USA 110 (50):20338–20343

Monk J, Nogales J, Palsson BO (2014) Optimizing genome-scale network reconstructions. Nat Biotechnol 32(5):447–452

Monk JM, Koza A, Campodonico MA, Machado D, Seoane JM, Palsson BO, Herrgård MJ, Feist AM (2016) Multi-Omics quantification of species variation of *Escherichia coli* links molecular features with strain phenotypes. Cell Syst 3(3):238–251. e12

Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, Takeuchi R et al (2017) iML1515, a knowledgebase that computes *Escherichia coli* traits. Nat Biotechnol 35(10):904–908

Moscatello N, Pfeifer BA (2018) Constraint-based metabolic targets for the improved production of heterologous compounds across molecular classification. AIChE J Am Inst Chem Eng 9 (July):293

Norsigian CJ, Kavvas E, Seif Y, Palsson BO, Monk JM (2018) iCN718, an updated and improved genome-scale metabolic network reconstruction of *Acinetobacter baumannii* AYE. Front Genet 9(April):121

O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BØ (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol 9(1):693

O'Brien EJ, Monk JM, Palsson BO (2015) Using genome-scale models to predict biological capabilities. Cell 161(5):971–987

Oberhardt MA, Puchałka J, Martins dos Santos VAP, Papin JA (2011) Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. PLoS Comput Biol 7(3): e1001116

Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28(3):245–248

Reed JL (2012) Shrinking the metabolic solution space using experimental datasets. PLoS Comput Biol 8(8):e1002662

Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nat Rev Genet 7(2):130–141

Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N (2016) Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods 13(5):435–438

Seif Y, Kavvas E, Lachance J-C, Yurkovich JT, Nuccio S-P, Fang X, Catoiu E, Raffatellu M, Palsson BO, Monk JM (2018) Genome-scale metabolic reconstructions of multiple salmonella strains reveal Serovar-specific metabolic traits. Nat Commun 9(1):3771

Shoaie S, Karlsson F, Mardinoglu A, Nookaew I, Bordel S, Nielsen J (2013) Understanding the interactions between bacteria in the human gut through metabolic modeling. Sci Rep 3:2532

Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot E, de Wouters T et al (2015) Quantifying diet-induced metabolic changes of the human gut microbiome. Cell Metab 22(2):320–331

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. Proc Natl Acad Sci USA 102(39):13950–13955

Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5(1):93–121

Thiele I, Jamshidi N, Fleming RMT, Palsson BØ (2009) Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. PLoS Comput Biol 5(3):e1000312

Bacterial Epigenomics: Epigenetics in the Age of Population Genomics

Poyin Chen, D. J. Darwin Bandoy, and Bart C. Weimer

Abstract Genome methylation in bacteria is an area of intense interest because it has broad implications for bacteriophage resistance, replication, genomic diversity via replication fidelity, response to stress, gene expression regulation, and virulence. Increasing interest in bacterial DNA modification is coming about with investigation of host/microbe interactions and the microbiome association and coevolution with the host organism. Since the recognition of DNA methylation being important in *Escherichia coli* and bacteriophage resistance using restriction/modification systems, more than 43,600 restriction enzymes have been cataloged in more than 3600 different bacteria. While DNA sequencing methods have made great advances there is a dearth of method advances to examine these modifications in situ. However, the large increase in whole genome sequences has led to advances in defining the modification status of single genomes as well as mining new restriction enzymes, methyltransferases, and modification motifs. These advances provide the basis for the study of pan-epigenomes, population-scale comparisons among pangenomes to link replication fidelity and methylation status along with mutational analysis of *mutLS*. Newer DNA sequencing methods that include SMRT and nanopore sequencing will aid the detection of DNA modifications on the ever-increasing whole genome and metagenome sequences that are being produced. As more sequences become available, larger analyses are being done to provide insight

P. Chen
School of Medicine, Harvard University, Boston, MA, USA

D. J. D. Bandoy
College of Veterinary Medicine, University of the Philippines Los Baños, Los Baños, Philippines

School of Veterinary Medicine, 100K Pathogen Genome Project, University of California Davis, Davis, CA, USA

B. C. Weimer (✉)
School of Veterinary Medicine, 100K Pathogen Genome Project, University of California Davis, Davis, CA, USA
e-mail: bcweimer@ucdavis.edu

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_11

into the role and guidance of bacterial DNA modification to bacterial survival and physiology.

Keywords Population bacterial genomics · Whole genome sequencing · Direct RNA sequencing · Gene expression

1 Introduction

Bacterial cellular functions are widely impacted via epigenetic modification, including bacteriophage infection, metabolism, virulence, persistence, replication, and genome plasticity. DNA modification in bacteria is of great interest because it is increasingly being linked to functional regulation processes in the organism and disease progression in mammals (Kumar and Rao 2013). DNA methylation was first recognized in *Escherichia coli* as part of restriction/modification systems (RMS) that limit and regulate bacteriophage infection. RMS are ubiquitous in the bacterial world with >43,600 RM recognized enzymes in >3600 bacteria (http://rebase.neb.com/rebase/rebase.html) (Roberts et al. 2010). Methylation primarily occurs at N^6adenine and C^5cytosine in many species, but only N^4cytosine is found in bacteria (Wion and Casadesus 2006; Kumar and Rao 2013). Recently, a new modification that regulates the redox status of the cell using DNA modification via a unique multifunctional alteration via phosphothioation was identified (Wang et al. 2019). Subsequently, DNA and RNA methylations were defined to play a central role in bacterial phenotypes that were not encoded in the genome but inherited in bacteria and do regulate gene expression in bacteria. Post-replication modification allows cells to rapidly adjust to local environmental conditions via gene expression changes that are not directly linked to genome variation yet require very dynamic shifts for survival and growth status.

An emerging area of investigation is the role of the microbiome on the host epigenome. Particular interest is paid to the role of the bacterial involvement in host cancer due to dysregulation of gene expression as cancer progresses. A comprehensive review of the state of progress that links infectious agents to cancer and host epigenome proposed that chronic inflammation was involved in the dysregulation of gene expression (Rajagopalan and Jha 2018). An intriguing hypothesis is that bacterial metabolism in utero can have long-lasting effect by regulating epigenetic modification of the maternal and fetal status in utero (Romano and Rey 2018). The complexity of the microbiome composition and metabolism leads one to expect a very complex system for the bacterial community to regulate the host epigenome. Farhana et al. (2018) reviewed the microbiome and its potential role in cancer. Of particular interest is that of *Helicobacter pylori* since it is associated with multiple states of disease in the progression from normal tissue to cancer with regional and human race differences since it has coevolved with humans for at least 80,000 years (Munoz-Ramirez et al. 2017), and it has a complex lifestyle in the microbial community within a unique location in the body that forces the organism to manage swings in pH, redox, and nutrient sources within minutes.

With the emergence of population genomics and metagenomics and large-scale whole-genome sequencing the vast amount of information has grown rapidly over a short time. With over 350,000 bacterial genomes in the public domain, a new challenge has grown in trying to conduct population epigenomes in bacteria and then associate those changes with change in the host to promote disease. Chen et al. (2014) described a method for population-scale approaches; however, more robust methods are now needed that include metagenome analysis as well.

Comparison of genomes using pangenomes and Big data approaches are progressing to link specific genes and alleles to disease. Population genomics is beginning to emerge (Weis et al. 2016) but it is disconnected to epigenomes and pangenome analysis at this point. Hence, focusing on specific genes and modifications is appropriate and providing results that can be linked to population genomics in the future.

2 Bacterial DNA Modifications and Biological Importance

On a biochemical level, epigenetic modification of the genome changes the accessibility of specific gene clusters and affinity of transcriptional regulators for their cognate promoters. This modulation of transcription accessibility and promoter affinity in turn translates to changes in bacterial response to environmental stimuli. Because epigenetic modifiers, such as RM systems and specific methyltransferases (MTases) themselves, are encoded on the chromosome as well as on plasmids, these elements can be transmitted vertically as a result of replication as well as horizontally as a result of horizontal gene transfer either via conjugation or phage. As mentioned above, DNA modification systems serve to identify and eliminate foreign DNA, but these DNA modifications also serve important roles in cell cycle progression, DNA repair, and regulation of gene expression.

2.1 Bacterial Histone-Like Proteins

Like eukaryotic histones, bacterial histone-like proteins assist in compacting the chromosome into a nucleoid structure (Thanbichler et al. 2005). Histone-like proteins can be classified into four different categories: histone-like proteins (HU), histone-like nucleoid structuring proteins (H-NS), integration host factors (IHF), and factors for inversion stimulation (FIS), further reviewed in Dorman and Deighan (2003) and Anuchin et al. (2011). To accomplish this task, bacteria utilize histone-like proteins to organize their DNA to minimize space utilization but also to regulate the expression of their DNA. These proteins work in a concerted manner to bind DNA and facilitate supercoiling into a nucleoid structure and regulate gene expression, these mechanisms were extensively reviewed previously (Dorman and Deighan 2003; Thanbichler et al. 2005; Dorman 2013; Takahashi 2014; Grainger 2016). Throughout the cell cycle, different histone-like proteins peak in concentration to regulate genes sets responsible for the progression of an actively replicating cell to a stationary phase cell, indicating that each one plays a unique role during specific stages of growth. Cycling histone-like proteins indicates that the pan-epigenome changes at different phases of growth. In addition to being related to different growth phases, expression of specific histone-like proteins is also induced in response to environmental stresses. The ability of environmental stimuli to change histone association with DNA suggests that pan-epigenetic shifts occur when an organism adapts to its environment. Examples are evident in the existence of microbes adapted to live in extreme environments as well as pathogens, such as *Brucella*, that are specifically adapted to live in their host. While these microbes no longer possess genes found in related species, it was epigenetic selection that led to the refinement of these genomes. Sustained pan-epigenetic shifts result in perpetually inactivated genes that are subsequently lost in future generations, resulting in differentiation between DNA modification and genotypes.

Although DNA methylation is frequently associated with RM systems and bacterial “immunity” against sources of foreign DNA, we are just beginning to understand the global impacts of DNA methylation on transcriptional regulation of gene expression. In addition to protein–DNA interactions affected by methylation, DNA modifications also regulate bacterial histone-like protein binding to DNA.

While MTases may indirectly impact gene expression through modulating histone-like protein–DNA interactions, MTases directly influence gene expression through the presence of recognition motifs located in promoter regions and protein-binding sites of genes. The methylation state of these regions work by modulating the affinity of RNA polymerase and transcriptional regulators such as leucine-responsive repeat protein (Lrp) and catabolite activator protein (Cap) to specific genes, among which include *dnaA*, *ppiA*, *yhiP*, and the *pap* operon (Tavazoie and Church 1998; Marinus and Casadesus 2009).

RM systems play a major role in bacterial immunity against foreign DNA. Another component of the bacterial “immune system” was recently discovered, termed clustered regularly interspaced palindromic repeats/CRISPR-associated (CRISPR/Cas). CRISPR systems are detectable in 1126 of the 2480 genomes analyzed to date (Grissa et al. 2007). Similar to phase variable regions of the genome, CRISPR/Cas systems are composed of short, conserved, DNA repeat sequences interspersed by stretches of variable sequences with *cas* genes adjacent to these regions. CRISPR/Cas systems recognize foreign nucleic acids, targeting them for degradation via RNA interference effector complexes composed of Cas proteins and CRISPR RNAs (Gasiunas et al. 2013). Though no associations between MTases and CRISPR/Cas have been proven, Hernández-Lucas et al. determined that *Salmonella* Typhi *casA* is under H-NS and Lrp regulation (Medina-Aparicio et al. 2011). In addition to immunity, CRISPR/Cas systems are also hypothesized to affect DNA mismatch repair with *E. coli* Cas1 involved in DNA segregation and mismatch repair (Babu et al. 2011; Westra et al. 2012). MTases and CRISPRs both share a number of common interacting partners involved in transcriptional regulation including Lrp and H-NS. While much remains to be learned about additional cellular roles of these systems, it is not improbable to expect a synergistic interaction in orchestrating essential cell processes.

2.2 DNA Modifications

Bacteria encode numerous restriction-modification (RM) systems that can be categorized into four main types. RM systems include the restriction endonuclease (REase), methyltransferase (MTase), and the specificity protein which facilitate targeted RM enzymatic activity to specific regions of DNA. RM systems require a specific unit, which enables RM targeting to a DNA recognition domain, a methyltransferase that modifies DNA with a methyl group, and an endonuclease that cleaves DNA (REase) with four types of RM systems described to date and catalogued in Rebase (Roberts et al. 2010). Briefly, Type I is characterized by an oligomeric MTase and REase complex with restriction occurring at variable distances from the recognition site. As the largest category with over 16,000 MTases identified, Type II system fall into numerous subcategories and are composed of either discreet or fused, MTase and REase subunits that cleave at or near the recognition site. Type III system cut at a fixed site away from the recognition sequence with the restriction enzyme activity contingent on association with the cognate MTase. Like Type I, Type IV system cleave at a variable distance from the recognition site but unlike the other three systems, the Type IV system is able to recognize and cleave hydroxymethylated and phosphorothioated DNA in addition to methylated DNA (Vasu and Nagaraja 2013; Loenen et al. 2014).

Originally discovered as a protective mechanism against bacteriophage infection, MTases selectively transfer the methyl group from SAM to the nitrogen atoms at position 4 in cytosine and position 6 of adenine (m^4C, m^6A) or the fifth carbon of cytosine (m^5C) within specific sequence motifs along the bacterial genome identified by the RM system recognition domain (Wilson 1991). These methylated sequences are resistant to endonuclease digestion by the restriction enzyme and are recognized by the RM system as a means of establishing self from nonself. Any phage DNA entering the host is assessed by the RM system and digested by the RM endonuclease if methylation is not detected by the corresponding recognition domain. To circumvent host restriction of phage DNA, bacteriophage often introduces their own MTases during infection. Due to the nature of RM enzyme–DNA dynamics, these MTases are often retained by the host following bacteriophage infection and transferred to subsequent generations, giving rise to orphan MTases lacking a reciprocal restriction enzyme (Labrie et al. 2010; Murphy et al. 2013).

Early experiments involving manipulation of RM systems produced viable cells with r + m + and r-m + phenotypes. Interestingly, an r + m- phenotype was lethal, suggesting that in the absence of DNA methylation, restriction enzymes will digest self-DNA, resulting in cell death (Arber 1965). In studying postsegregational killing by RM systems, Kobayashi et al. observed a larger amount of MTases molecules relative to REase in steady-state cells. However, dysregulation of cellular MTase and REase levels led to increased cell death due to Res-induced double-strand breaks in the chromosome (Ichige and Kobayashi 2005). These results further highlight a characteristic true of all RM systems in which MTases are fully functional without the cognate restriction enzyme; however, the restriction enzyme activity is contingent on the presence of the MTase. Easy acquisition and retention of foreign MTases—termed orphan MTases—by host bacteria contributes to the increased diversity of MTases in relation to restriction enzymes with possible methyltransferase sources being mobile elements acquired through transduction or mating events (Murphy et al. 2013).

DNA Adenine Methylation DAM DNA adenine methylation (Dam) is the predominant methylation found in bacteria and is accomplished by bacterial methyltransferases (MTases). Dam MTases are widespread throughout all genera of bacteria, with some MTases sharing the same recognition motif and other MTase recognition sites being species, if not strain, specific. The presence of hydrophobic methyl groups either on both strands of DNA (fully methylated) or a single strand of DNA (hemi methylated) serve to modulate gene expression by way of modulating the affinity of DNA-binding proteins for specific regions of DNA.

Survival in a niche environment such as the human body requires careful and concerted regulation of numerous genes, ranging from stress response and nutrient acquisition to manipulation of host processes in the case of pathogenic bacteria. Although bacterial pathogens have coevolved with their hosts (Hongoh et al. 2005), the standard transmission cycle of some pathogens dictate that they may spend some time outside of their human host and in environments that are suboptimal in moisture and nutrients but can contain antimicrobial compounds (Harb et al. 2000). Transitioning from an environmental lifestyle to a host-adapted lifestyle requires a large shift in the gene expression and protein profile of a pathogen. With the magnitude of gene regulation needed to facilitate this lifestyle change, it is reasonable to consider the role of epigenetics in driving these changes (Low et al. 2001).

E. coli The *pap* operon of *E. coli* encodes the pyelonephritis-associated pilus. While *pap* is under methylation-mediated transcriptional control, Pap expression is also regulated by methylation-mediated phase variation. Mechanistically, Dam competes with transcriptional regulators, such as Lrp, a global transcriptional activator, for access to recognition domains wherein methylation of the domain determines the pilus ON/OFF state (Casadesus and Low 2006). Similar mechanisms governing pilus formation and phase variation are also documented in many other bacteria including *Salmonella, S. aureus*, *H. influenza, Neisseria,* and *H. pylori* (Srikhanta et al. 2005, 2011).

Salmonella This organism is broadly modified (Table 1) over the genome with specific motifs. Within the same *Salmonella* virulence plasmid, H-NS represses *finP* in a Dam-dependent manner while repressing *traJ* in a Dam-independent manner. These observations bring to light the impact of structural differences in nucleoids of *dam* + vs *dam-* genomes and the outcome of these structural differences on gene expression (Marinus and Casadesus 2009). In addition to histone-like

Table 1 Epigenetic modification of selected *Salmonella* serotypes determined using SMRT sequencing (Weimer, unpublished)

	Bareilly (SAL2881)	Heidelberg (CFSAN000318_04)	Javiana (CFSAN001992_73)	Typhimurium (CFSAN001921_01)	St Paul (SP3)
5′-G**A**TC-3′/3′-CT**A**G-5′	🟩	🟩	🟩	🟩	🟩
5′-CAG**A**G-3′/3′-GTCTC-5′	🟩	🟩	🟩	🟩	🟩
5′-ATGC**A**T-3′/3′-T**A**CGTA-5′	▨	▨	▨	▨	
5′-CAG**C**TG-3′/3′-GT**C**GAC-5′	🟩				
5′-GATC**A**G-3′/3′-CTAGTC-5′				🟩	
5′-ACC**A**NCC-3′/3′-TGGTNGG-5′		🟩			
5′-CCG**A**N5GTC-3′/3′-GGCTN5C**A**G-5′	🟩				
5′-G**A**GN6RTAYG-3′/3′-CTCN6Y**A**TRC-5′		🟩		🟩	🟩
5′-GN2T**A**YN5RTGG-3′/3′-CN2ATRN5Y**A**CC-5′			🟩		
5′-G$_{ps}$AAC-3′/3′-CTT$_{ps}$G-5′					▨

Increasing shades of green indicate higher modification in each isolate with the most being 100% and the least being 10%. No shade indicates no modification and bold base indicates location of modification

proteins, DNA methylation, specifically adenine methylation (Dam) is known to be involved in regulating host colonization. PhoP, a master regulator of *Salmonella* virulence, binds DNA in a dam-dependent manner (Heithoff et al. 1999). Deletion or over expression of an MTase results in whole genome-wide change in transcription profiles. While *Salmonella* Typhimurium Dam mutants do not exhibit growth-related deficiencies, Dam-deficient *Salmonella* exhibits a 10,000-fold increase in the lethal dose required to kill 50% of a mouse population (LD_{50}) (Low et al. 2001). Transcriptional profiling of Dam-deficient *Salmonella* attributes attenuation to an induction of *spvB*, along with over 35 other infection-associated genes and a reduction in *sipABC* transcripts (Garcia-Del Portillo et al. 1999).

The amount of information in specific organisms that have a minor role in disease or lack a large amount of whole genome sequence has very little pan-epigenome information. Chen et al. (2017) examined the epigenome of *L. monocytogenes* (Table 2) to find a complex pattern of modification that was not observed to be associated with pathogenicity. Virulence genes were heavily methylated, but no observable pattern emerged to uncover how methylation was involved in virulence.

DNA Cytosine Methylation (DCM) Unlike adenine methylation that has been functionally characterized in numerous bacterial systems, DNA cytosine methylation (Dcm) remains relatively understudied. Best characterized in *E. coli*, Dcm appears to confer resistance against restriction by the REase, EcoRII (Bigger et al. 1973; Boye and Lobner-Olesen 1990). Functionally, Dcm acts as an antitoxin against EcoRII restriction. Because Dcm and EcoRII share the same recognition

Table 2 Epigenome prevalence of modification in *Listeria monocytogenes* isolates involved in a foodborne illness outbreak derived from pathogenesis association (Chen et al. 2017)

	Serotype	1/2a								1/2b					4b		
Methyltransferase Specificity	**Modified Base**	**861**	**878**	**899**	**1846**	**2074**	**2625**	**2626**	**2676**	**859**	**867**	**911**	**2624**	**G4599**	**1493**	**1494**	**1495**
5′-G**A**TC-3′ 3′-CT**A**G-5′	m6A									99.5		98.9					
5′-GA**T**C-3′ 3′-C**T**AG-5′	X		8.4			12.2	15.8										
5′-G**A**CN$_5$GGT-3′ 3′-CTGN$_5$CC**A**-5′	m6A												98.9 98.9				
5′-G**A**N$_6$TGCG-3′ 3′-CTN$_6$**A**CGC-5′	m6A				99.7 99.6			100 99.8	99.8 99.9								
5′-T**A**CBN$_6$GTNG-3′ 3′-ATGVN$_6$C**A**NC-5′	m6A													99.7 99.8			
5′-TAGR**A**G-3′ 3′-ATCYTC-5′	m6A											99.3					
5′-GT**A**TCC-3′ 3′-CAT**A**GG-5′	m6A														99.6 99.7	99.1 98.8	99.2 98.0

Bold letters indicate the modified base. Numbers indicate the percentage of that motif modified in the genome using SMRT sequencing. Boxes with two sets of numbers indicates the strand specific prevalance methylation

sequence—C^{m}CWGG—Dcm is able to methylate sites that would otherwise be targeted for EcorII restriction (Palmer and Marinus 1994). In this manner, Dcm serves a protective function against a parasitic RM system (Takahashi et al. 2002). Dcm is also associated with mobile element rearrangements in the *E. coli* genome involving bacteriophage lambda recombination and TN3 transposition (Korba and Hays 1982; Yang et al. 1989). On a whole genome level, evidence suggests that Dcm is involved in transcriptional and translational regulation of ribosome activity to decrease the expression of ribosomal proteins during stationary phase (Militello et al. 2012).

Phosphorothioate Modification A third, recently discovered DNA modification that naturally occurs in bacteria is phosphorothioate (PT) modification wherein the oxygen atom in a phosphate moiety of the DNA backbone is replaced by sulfur (Eckstein 2014). The ability to carry out PT modifications is contingent on the presence of the *dnd* gene clusters, *dndABCDE*, the modification component, and *dndFGH*, the restriction component although their presence can be mutually exclusive (Tong et al. 2018). First discovered in *Streptomyces lividans*, informatics analyses of *dnd* gene clusters has since revealed a wide distribution of PT modifications in bacterial genomes (He et al. 2007; Wang et al. 2011, 2019). Abrogation of PT modifications led to increased double-stranded DNA breaks in *Salmonella* and oxidative stress due to significant metabolic changes in *Pseudomonas fluorescens* (Cao et al. 2014; Gan et al. 2014; Tong et al. 2018).

Undiscovered Modifications Next-generation sequencing techniques that incorporate measurement of polymerase kinetics can detect structural differences to individual nucleotides that would otherwise have been overlooked (Rhoads and Au 2015). By comparing the pattern of polymerase kinetics to previously characterized patterns, we can informatically identify DNA modifications at the single nucleotide level and characterize epigenetic patterns on the whole genome level (Schadt et al. 2013). The use of this technology in whole genome sequencing has also recorded polymerase kinetics patterns that are not yet associated with a known DNA modification (Chen et al. 2017). These data suggest that there is unprecedented diversity to epigenetic modifications that we have yet to uncover. Epigenetic modifications that have been characterized thus far are responsible for numerous physiological processes including defense against foreign DNA, gene regulation, and DNA replication and mismatch repair. The implications of uncharacterized modifications on epigenomic regulation potentially have far-reaching implications for interactions within a niche and interaction with the host for survival and persistence. As additional advances are made in next-generation sequencing and RNAseq, it may be possible to define methylation directly in situ, which is a current limitation.

2.3 DNA Replication and Chromosome Sorting

Bacteria encode proteins near their chromosomal origin of replication (*oriC*) that facilitate the timing of replication initiation and help to carry out the chromosome segregation during replication (Ogden et al. 1988; Boye and Lobner-Olesen 1990; Campbell and Kleckner 1990). Due to the time-sensitive nature of replication initiation, DNA replication-associated protein levels must be tightly coordinated with cellular replicative machinery. To accomplish this task, bacteria encode a higher density of GATC methylation sites around the origin of replication and utilize DNA methylation to modulate the affinity of replication-associated proteins to DNA. Methylation around *oriC* regulates the recruitment of replication initiation proteins including the initiator of replication, DnaA. Furthermore, GATC methylation motifs also exist in the promoter region of *dnaA*, allowing for transcriptional regulation of replication (Campbell and Kleckner 1990). During DNA replication, both copies of the chromosome must be accurately sorted into the corresponding cell. After replication, DNA is in a hemi-methylated state. Methylation at *oriC* sequesters the origin replication initiation and prevents reinitiation of DNA replication. Additionally, global hemi-methylation of newly replicated DNA facilitates chromosome binding to designated areas of the cell membrane such that individual chromosomes may be accurately partitioned into each daughter cell (Ogden et al. 1988).

2.4 Mismatch Repair and Evolution

Bacterial DNA polymerases are capable to replicating DNA with high fidelity, but replication errors still arise at a rate of 10^{-9} to 10^{-11} errors per base pair (Drake et al. 1998). When these replication errors arise, the cell must have a way of identifying the correct template with which to correct the mistake. Template and newly replicated strands of DNA are differentially methylated to differentiate from one another with the template being methylated and the newly replicated strand remaining unmethylated. First described in *Streptococcus pneumoniae* and further characterized in *E. coli*, this methyl-directed mismatch repair system was identified as MutHLS (Glickman and Radman 1980; Claverys and Lacks 1986) (Fig. 1). MutS binds to mismatched base pairs while the methyl-sensitive endonuclease MutH nicks the DNA at the mismatched site. MutL recruits the DNA repair machinery to correct the mismatch. Both the loss of MTases and overexpression of MTases are correlated with deficient mismatch repair due to a dysregulation between methylation and DNA replication kinetics. In *dam* mutants, the inability to methylate the template strand leads to inaccurate mismatch repair and vertical transmission of mutations arising from DNA replication. *Dam* mutants are unable to methylate the template strands of replicated DNA, leading *MutHLS* inability to identify the strand of DNA containing the mutation for mismatch repair. In this regard, the pan-epigenome directly influences the accumulation of SNPs that arise during replication. Due to the mobile nature of RMS systems, over time the loss or acquisition of additional MTase systems may influence the global methylation status of a genome.

3 Epigenetic Detection Methods and Approaches

Nucleotide modification by methylation is a prevalent feature in living organisms. In bacteria, base methylation is a form of defense system against bacteriophage or foreign genetic material. The defense system works by detecting sequence motifs of nucleotides and cuts it using an endonuclease as a preemptive strike against foreign genome. Bacterial DNA is spared from the cutting with the action of the methylase. This is known as the restriction-modification system (RMS). Aside from defensive function, the restriction modification system also performs genomic regulatory functions in bacteria. Due to the huge impact of the restriction modification system in the lifestyle of bacteria with regard to pathogenicity, prokaryotic epigenomics is an emerging field primarily driven by recent technological advancement in sequencing capability. The transformational aspect is mainly on the scalability of methylation analysis at the genomic level. This has opened up doors for genome-wide methylation analysis.

What are the key considerations in doing large-scale high-throughput epigenomics research? Genome-wide methylation projects' considerations are determined by costs, ease of library construction and preparation, access to equipment or core

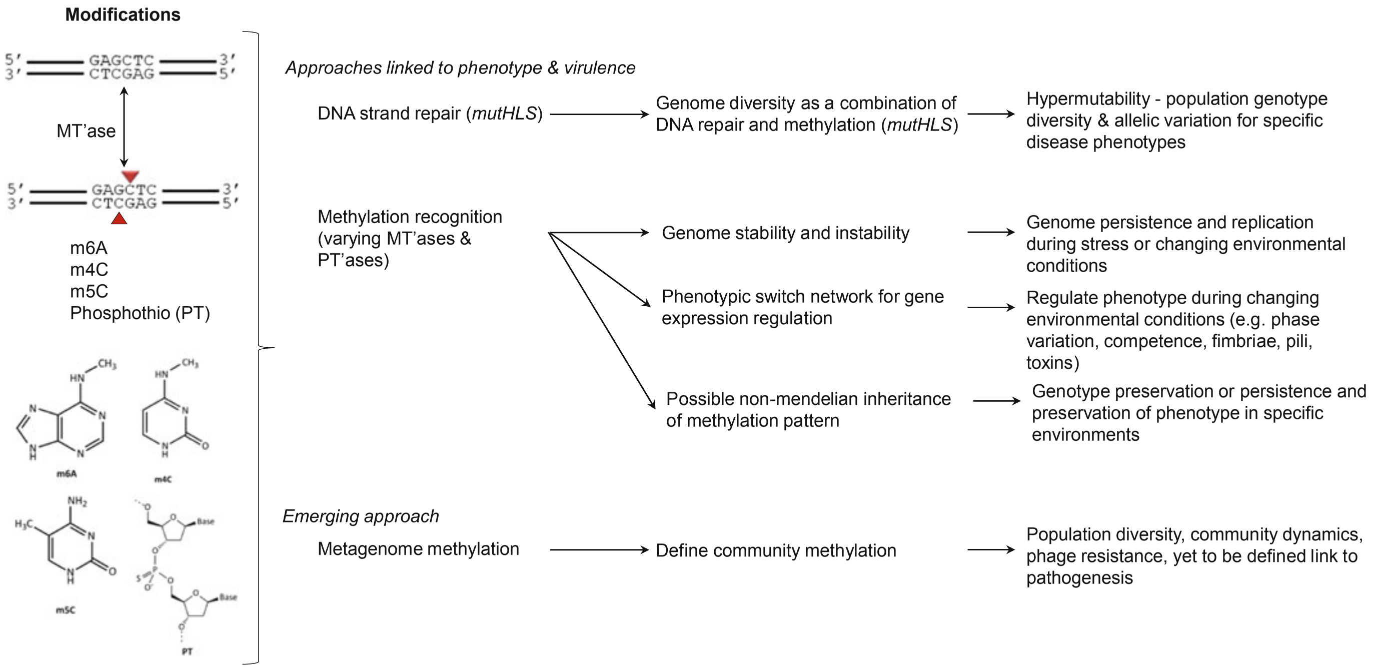

Modifications
5′ GAGCTC 3′
3′ CTCGAG 5′
MT'ase
5′ GAGCTC 3′
3′ CTCGAG 5′
m6A
m4C
m5C
Phosphothio (PT)
m6A
m4C
m5C
PT
Approaches linked to phenotype & virulence
DNA strand repair (*mutHLS*)
Genome diversity as a combination of DNA repair and methylation (*mutHLS*)
Hypermutability - population genotype diversity & allelic variation for specific disease phenotypes
Methylation recognition (varying MT'ases & PT'ases)
Genome stability and instability
Genome persistence and replication during stress or changing environmental conditions
Phenotypic switch network for gene expression regulation
Regulate phenotype during changing environmental conditions (e.g. phase variation, competence, fimbriae, pili, toxins)
Possible non-mendelian inheritance of methylation pattern
Genotype preservation or persistence and preservation of phenotype in specific environments
Emerging approach
Metagenome methylation
Define community methylation
Population diversity, community dynamics, phage resistance, yet to be defined link to pathogenesis

Fig. 1 DNA modifications found in bacteria and the associated implications for bacterial populations, phenotype variation, and host impact

facility, availability of suitable kits for library construction and downstream bioinformatic analysis. The level of resolution of epigenomic modification data from crude to precise distinguishes the possible technological options appropriate for the pipeline. The above-mentioned considerations as well as the underlying technology will be covered in the succeeding sections.

3.1 Pre-sequencing Methods for Genome Methylation: LC-MS, HPLC-UV, and ELISA

The pre-sequencing methods are generally used for basic research and their capability to quantify methylation at the genomic scale. While this ability to quantify methylation at the genome scale provides a big picture setting of methylation, mapping the methylation sites to the specific regions in the genome is not possible. The scalability for population-scale bacterial epigenomics is limited and hence has limited the applicability of these methods to a few niche research papers.

The key steps in the analytical workflows are DNA extraction, genomic fragmentation, enrichment, and quantification using chromatography or mass spectrometry. The options for genomic fragmentation are thermal, chemical, and enzymatic hydrolysis. The resulting digested DNA monomers is enriched using size-exclusion, liquid extraction, solid phase extraction, or preparative liquid chromatography. Analyte ions are separated by the mass-to-charge ratios in mass spectrometry, allowing binning of the DNA monomers (Tretyakova et al. 2013).

Genome wide methylation using analytical methods particularly HPLC-based methods have been recently described (Yotani et al. 2018). High-performance liquid chromatography-ultraviolet (HPLC-UV) enables quantification and identification by separating the different components. This is accomplished by pushing the components using pressurized liquid solvent through a column filled with solid adsorbent material. The differences between the materials result to variation in flow rates allowing separation of the components. In bacterial DNA methylation analysis, this method is applied to quantify the separated methylated and unmethylated deoxynucleosides.

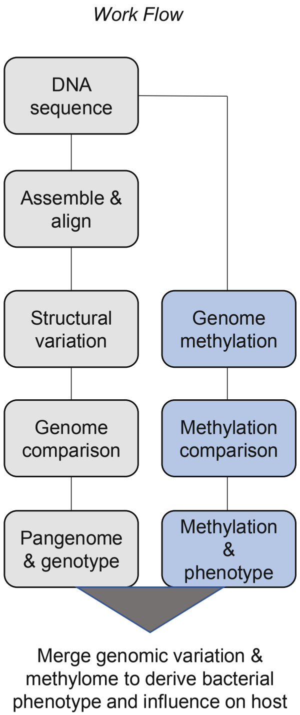
For crude global methylation analysis, numerous commercial ELISA (enzyme-linked immunosorbent assay) kits are available. The high level of variance is the primary reason for the lack of precision of ELISA kits in epigenomics, but the ease of use is sufficient to capture huge differences in methylation. The target DNA is immobilized on ELISA plate and specific primary antibody against methylated nucleoside is applied followed by a secondary antibody that can be detected using colorimetric methods.

The requirement for specialized equipment for LC-MS and HPLV-UV has restricted the use of the following methods for genome-wide methylation. While relative quantification is possible, mapping the methylation is not possible and hence population-scale analysis is not possible. The technical challenges of doing the work hinders its large-scale application.

3.2 Next-Generation Sequencing-Based Methods

The key shortcoming in using analytical methods for bacterial epigenomics is inability to identify methylation loci. This deficiency has predominantly filled by next-generation sequencing technology that can simultaneously capture sequence and methylation data (Fig. 2). The prevailing choice for combined sequencing and methylation platform is single molecule real-time (SMRT) sequencing by PacBio. Data is captured for ^{6}mA, ^{4}mC, and ^{5}mC parallel to sequencing data based on the kinetics of DNA synthesis reactions. This enables genome-wide mapping of methylated and unmethylated loci. Modified bases have not been a routinely included in the Sanger-based sequence analysis and has posed significant technological challenge until the arrival next-generation sequencing options. DNA treatment with bisulfite converts unmodified cytosine to uracil, enabling discrimination between modified and unmodified cytosine using various sequencing platform.

SMRT sequencing follows the typical workflow for next-generation sequencing with library construction after DNA extraction (Kong et al. 2017). The protocols for automated PacBio 10 kb library construction have been published, which can immensely improve efficiency of performing epigenomic research. A crucial requirement for successful high-throughput sequencing run is high molecular weight genomic DNA. Agilent 2200 TapeStation Nucleic Acid System has been used to determine the quantity and size distribution of purified genomic DNA (Kong et al. 2014) as well as the 260/280 and 260/230 ratio using Nanodrop 2000 UV–vis spectrophotometer (ThermoFisher Scientific, Waltham MA). The DNA integrity number (DIN) is a suitable tool for determining the quality of genomic DNA for further processing (Kong et al. 2016) and methods exist for automated construction of the sequencing library (Kong et al. 2017). The core basis for SMRT sequencing is based on restrictions of light illumination of immobilized target DNA and polymerase using zero-mode waveguide (Rhoads and Au 2015). Signal detection of the cleaved fluorescent dye from the nucleotide molecule is the basis for base calling. The bulk of the most technically challenging aspect of the analysis is within the post sequencing bioinformatic pipeline. DNA methylation detection and quantification analysis are done in PacBio SMRT analysis platform (http://www.pacb.com/devnet/code.html). After sequencing, raw reads are trimmed to remove adapter sequences and then aligned to a reference using BLASR (v1) (Chaisson and Tesler 2012). DNA methylated sites are then determined using kinetic analysis of the genomic alignment. MotifFinder clusters the methylated sites to motifs targeted by methylases. This platform also allows discovery of novel restriction-modification genes. Homology is inferred bioinformatically using databases like SeqWare for cloud applications (O'Connor et al. 2010).



Methods to obtain sequence and methylome

Bisulfite sequencing	SMRT sequencing	Nanopore sequencing	High throughput association mapping
5mC, (4mC)	6mA, (5mC), 4mC, PT	6mA, 5mC, 4mC, (PT)	Increasing number of WGS is ahead of methylome determination of biological importance of modification
Common approach & standard methods	Emerging approach that is limited by cost for wide-spread use to link complete genomes and methylomes	Emerging approach that has potential to derive real-time methylome determination	Uncommon approach that would benefit WGS and metagenome interrogation of functional importance of individual genes and methylation variation to function
Requires TET treatment and additional sequencing run to detect 4mC	Requires TET treatment or deep sequencing to detect 5mC	Methylation obtained directly from sequence run	Requires direct mapping and phenotype metadata
Cannot detect 6mA or PT modifications	Cost can be high with large samples; metagenome methylomes are emerging	Additional development of algorithms and integration methods in process	Not widely used but large need
Draft genome information using short read libraries	Long read sequence likely complete genome assemblies	Long read sequence likely complete genome assemblies	WGS and metagenome input data would be viable to link host impact and virulence
This technique is being used less in favor of long read technologies	Used widely with WGS but limited by cost; use with metagenomes emerging	Real time sequence that has potential to provide real time methylome; use with metagenomes and RNA directly emerging	Emerging need to utilize the ever increasing public WGS and metagenomes

Fig. 2 DNA sequencing approaches to determine the methylome using next-generation workflows and comparison of output from each method. Modifications with brackets indicate additional chemistry to be done to determine the specific modification

The development in sequencing technology allowed large-scale analysis of prokaryotes (Blow et al. 2016). Base resolution methylation was captured in unprecedented detail and scale using SMRT sequencing initially. The variety of methylation was found on about 800 different loci in this study, indicative of precise specificities of methylation present in the bacterial organism. With the use of SMRT sequencing, the methylation repertoire was significantly increased. This highlights the key advantage of SMRT sequencing to further enhanced the recognition specificities of the methylase. Novel mechanistic epigenomic findings include: Type I RM system cleavage of DNA at large distances from their recognition sites, while both Type II and Type III systems incomplete cleavage pattern. This epigenomic feature is problematic for digestion-based analytical methods. The predilection of these RMS is toward m4C and m6A, which are readily detected by SMRT sequencing. Another understudied aspect of methylation is the orphaned methylases, which are common in prokaryotes. This relatively understudied group includes 100 Type II methylases. One novel discovery is potential regulatory control due to the genomic pattern associated with the orphan methylases which are located on noncoding sequences upstream of genes. This potential regulatory role was is widely distributed across the prokaryotic organism. In another study, a deeper resolution analysis such as identification and quantification of methylation motifs, correlation with methylases of methylation motifs using REBASE (Roberts et al. 2015) and identification of orphaned methylases has been done in large scale in organisms like *Listeria* (Chen et al. 2017). This study reported lineage- and clade-specific patterns of restriction-modification system (RMS). Type II RMS dominates with its presence in 256 out of 302 genomes, followed by Type I with 110 genomes, Type IV with 73 and lastly by Type III with 25 genomes. Methylation motifs were also described. These studies highlight the large-scale applicability of sequencing-based epigenomic study to unravel population-scale dynamics and patterns.

On a mechanistic level using fine-scale analysis, Fang et al. explored 6 mA methylation in a Shiga toxin-producing a strain of *E. coli* 0104:H4 Germany outbreak isolate predicted to produce 10 methylases that result in the 6-mA modification (Fang et al. 2012). A phage-encoded modification system capable of targeting hundreds of loci within the *E. coli* 0104:H4 isolate. This discovery of phage-encoded modification system-associated virulence had no prior examples in *E. coli*, illustrating the immense power to untangle epigenomic clues using sequencing platforms.

4 Conclusion and Future Direction

The epigenomic studies relied heavily on bioinformatics to deduce motifs that were highly enriched by modification with specific methylases. These studies discovered novel methylase specificities, quantified methylation activity, identified novel enzyme activity, which targets only one strand of DNA and promiscuous gene lacking specificity. Such precision is only possible with sequencing technology coupled with methylation detection capability. As sequencing technologies advance, the definition of modification will become increasingly important in biological function interpretation. A current limitation is that the vast amount of whole genome sequence and the limited number of methods to locate and estimate the modifications. A proxy for this limitation is to examine the RMS enzymes, which is interesting, but not direct enough to derive biologically accurate information. This method also suffers from informatics methods that can be applied on a comparative population scale, as can be done with pangenomes, but not pan-methylomes for bacteria. MethBank is available for a few mammals and plants (Li et al. 2018). The rate of bacterial genome production is only increasing. As such, a need exists to interrogate methylome of the organism at the speed of sequencing. This is not available and is a severe limitation in understanding bacterial growth, survival, and association; which is also true of metagenome interrogation as well. A great step forward would be to have a similar database for bacteria with the ability to allow pangenome and pan-methylome comparisons.

The field is poised to link the bacterial methylation status with the host methylation composition as it relates to disease. However, the dynamic nature of the microbiome, gene expression, and methylation in the bacterial component is a substantial challenge. Initial stages of examining the microbiome sequence for RMS enzymes are a starting point that will aid in understanding the complement of modifications that are possible. The beginning of this work has started in cancer progression and to some degree single organisms, such as *H. pylori*, in the development of various stages of cancer progression.

Bacterial metagenome production will increase with the expanded use of real-time sequencing technologies, such as nanopores. However, limitations in analysis and the dynamic nature of the bacterial DNA modification must be addressed to make substantial progress in linking it to phenotype. Future prospects of examining methylation are very exciting and there are many needs in the bioinformatic comparative analysis, especially in pathogens associated with chronic diseases.

References

Anuchin AM, Goncharenko AV, Demidenok OI, Kaprel'iants AS (2011) Histone-like proteins of bacteria (review). Prikl Biokhim Mikrobiol 47(6):635–641

Arber W (1965) Host-controlled modification of bacteriophage. Annu Rev Microbiol 19:365–378

Babu M, Beloglazova N, Flick R, Graham C, Skarina T, Nocek B, Gagarinova A, Pogoutse O, Brown G, Binkowski A, Phanse S, Joachimiak A, Koonin EV, Savchenko A, Emili A, Greenblatt J, Edwards AM, Yakunin AF (2011) A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Mol Microbiol 79(2):484–502

Bigger CH, Murray K, Murray NE (1973) Recognition sequence of a restriction enzyme. Nat New Biol 244(131):7–10

Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, Froula J, Kang DD, Malmstrom RR, Morgan RD, Posfai J, Singh K, Visel A, Wetmore K, Zhao Z, Rubin EM, Korlach J, Pennacchio LA, Roberts RJ (2016) The Epigenomic landscape of prokaryotes. PLoS Genet 12(2):e1005854

Boye E, Lobner-Olesen A (1990) The role of dam methyltransferase in the control of DNA replication in *E. coli*. Cell 62(5):981–989

Campbell JL, Kleckner N (1990) *E. coli* oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. Cell 62 (5):967–979

Cao B, Cheng Q, Gu C, Yao F, DeMott MS, Zheng X, Deng Z, Dedon PC, You D (2014) Pathological phenotypes and in vivo DNA cleavage by unrestrained activity of a phosphorothioate-based restriction system in salmonella. Mol Microbiol 93(4):776–785

Casadesus J, Low D (2006) Epigenetic gene regulation in the bacterial world. Microbiol Mol Biol Rev 70(3):830–856

Chaisson MJ, Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13:238

Chen P, Jeannotte R, Weimer BC (2014) Exploring bacterial epigenomics in the next-generation sequencing era: a new approach for an emerging frontier. Trends Microbiol 22(5):292–300

Chen P, den Bakker HC, Korlach J, Kong N, Storey DB, Paxinos EE, Ashby M, Clark T, Luong K, Wiedmann M, Weimer BC (2017) Comparative genomics reveals the diversity of restriction-modification systems and DNA methylation sites in *Listeria monocytogenes*. Appl Environ Microbiol 83(3)

Claverys JP, Lacks SA (1986) Heteroduplex deoxyribonucleic acid base mismatch repair in bacteria. Microbiol Rev 50(2):133–165

Dorman CJ (2013) Co-operative roles for DNA supercoiling and nucleoid-associated proteins in the regulation of bacterial transcription. Biochem Soc Trans 41(2):542–547

Dorman CJ, Deighan P (2003) Regulation of gene expression by histone-like proteins in bacteria. Curr Opin Genet Dev 13(2):179–184

Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148(4):1667–1686

Eckstein F (2014) Phosphorothioates, essential components of therapeutic oligonucleotides. Nucleic Acid Ther 24(6):374–387

Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado OJ, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Keren-Paz A, Chess A, Roberts RJ, Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE (2012) Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. Nat Biotechnol 30(12):1232–1239

Farhana L, Banerjee HN, Verma M, Majumdar APN (2018) Role of microbiome in carcinogenesis process and epigenetic regulation of colorectal cancer. Methods Mol Biol 1856:35–55

Gan R, Wu X, He W, Liu Z, Wu S, Chen C, Chen S, Xiang Q, Deng Z, Liang D, Chen S, Wang L (2014) DNA phosphorothioate modifications influence the global transcriptional response and protect DNA from double-stranded breaks. Sci Rep 4:6642

Garcia-Del Portillo F, Pucciarelli MG, Casadesus J (1999) DNA adenine methylase mutants of salmonella typhimurium show defects in protein secretion, cell invasion, and M cell cytotoxicity. Proc Natl Acad Sci USA 96(20):11578–11583

Gasiunas G, Sinkunas T, Siksnys V (2013) Molecular mechanisms of CRISPR-mediated microbial immunity. Cell Mol Life Sci 71(3):449–465

Glickman BW, Radman M (1980) *Escherichia coli* mutator mutants deficient in methylation-instructed DNA mismatch correction. Proc Natl Acad Sci USA 77(2):1063–1067

Grainger DC (2016) Structure and function of bacterial H-NS protein. Biochem Soc Trans 44 (6):1561–1569

Grissa I, Vergnaud G, Pourcel C (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8:172

Harb OS, Gao LY, Abu Kwaik Y (2000) From protozoa to mammalian cells: a new paradigm in the life cycle of intracellular bacterial pathogens. Environ Microbiol 2(3):251–265

He X, Ou HY, Yu Q, Zhou X, Wu J, Liang J, Zhang W, Rajakumar K, Deng Z (2007) Analysis of a genomic island housing genes for DNA S-modification system in Streptomyces lividans 66 and its counterparts in other distantly related bacteria. Mol Microbiol 65(4):1034–1048

Heithoff DM, Sinsheimer RL, Low DA, Mahan MJ (1999) An essential role for DNA adenine methylation in bacterial virulence. Science 284(5416):967–970

Hongoh Y, Deevong P, Inoue T, Moriya S, Trakulnaleamsai S, Ohkuma M, Vongkaluang C, Noparatnaraporn N, Kudo T (2005) Intra- and interspecific comparisons of bacterial diversity and community structure support coevolution of gut microbiota and termite host. Appl Environ Microbiol 71(11):6590–6599

Ichige A, Kobayashi I (2005) Stability of EcoRI restriction-modification enzymes in vivo differentiates the EcoRI restriction-modification system from other postsegregational cell killing systems. J Bacteriol 187(19):6612–6621

Kong N, Ng W, Azarene F, Carol Huang B, Kelly L, Weimer BC (2014) Quality control of library construction pipeline for PacBio SMRTbell 10kb library using Agilent 2200 TapeStation. Agilent Technologies, Santa Clara, CA. https://doi.org/10.13140/RG.2.1.4339.4644

Kong N, Ng W, Cai L, Leonardo A, Weimer BC (2016) Integrating the DNA integrity number (DIN) to assess genomic DNA (gDNA) quality control using the Agilent 2200 TapeStation system. Agilent Technologies, Santa, Clara, CA, pp 1–6

Kong N, Ng W, Thao K, Agulto R, Weis A, Kim KS, Korlach J, Hickey L, Kelly L, Lappin S, Weimer BC (2017) Automation of PacBio SMRTbell NGS library preparation for bacterial genome sequencing. Stand Genomic Sci 12:27

Korba BE, Hays JB (1982) Partially deficient methylation of cytosine in DNA at CCATGG sites stimulates genetic recombination of bacteriophage lambda. Cell 28(3):531–541

Kumar R, Rao DN (2013) Role of DNA methyltransferases in epigenetic regulation in bacteria. In: Kundu TK (ed) Epigenetics: development and disease. Springer, Dordrecht, pp 81–102

Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. Nat Rev Microbiol 8(5):317–327

Li R, Liang F, Li M, Zou D, Sun S, Zhao Y, Zhao W, Bao Y, Xiao J, Zhang Z (2018) MethBank 3.0: a database of DNA methylomes across a variety of species. Nucleic Acids Res 46(D1):D288–D295

Loenen WA, Dryden DT, Raleigh EA, Wilson GG, Murray NE (2014) Highlights of the DNA cutters: a short history of the restriction enzymes. Nucleic Acids Res 42(1):3–19

Low DA, Weyand NJ, Mahan MJ (2001) Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. Infect Immun 69(12):7197–7204

Marinus MG, Casadesus J (2009) Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. FEMS Microbiol Rev 33(3):488–503

Medina-Aparicio L, Rebollar-Flores JE, Gallego-Hernandez AL, Vazquez A, Olvera L, Gutierrez-Rios RM, Calva E, Hernandez-Lucas I (2011) The CRISPR/Cas immune system is an operon regulated by LeuO, H-NS, and leucine-responsive regulatory protein in *Salmonella enterica* serovar Typhi. J Bacteriol 193(10):2396–2407

Militello KT, Simon RD, Qureshi M, Maines R, VanHorne ML, Hennick SM, Jayakar SK, Pounder S (2012) Conservation of Dcm-mediated cytosine DNA methylation in *Escherichia coli*. FEMS Microbiol Lett 328(1):78–85

Munoz-Ramirez ZY, Mendez-Tenorio A, Kato I, Bravo MM, Rizzato C, Thorell K, Torres R, Aviles-Jimenez F, Camorlinga M, Canzian F, Torres J (2017) Whole genome sequence and phylogenetic analysis show helicobacter pylori strains from Latin America have followed a unique evolution pathway. Front Cell Infect Microbiol 7:50

Murphy J, Mahony J, Ainsworth S, Nauta A, van Sinderen D (2013) Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. Appl Environ Microbiol 79(24):7547–7555

O'Connor BD, Merriman B, Nelson SF (2010) SeqWare query engine: storing and searching sequence data in the cloud. BMC Bioinformatics 11(Suppl 12):S2

Ogden GB, Pratt MJ, Schaechter M (1988) The replicative origin of the *E. coli* chromosome binds to cell membranes only when hemimethylated. Cell 54(1):127–135

Palmer BR, Marinus MG (1994) The dam and dcm strains of *Escherichia coli*—a review. Gene 143 (1):1–12

Rajagopalan D, Jha S (2018) An epi(c)genetic war: pathogens, cancer and human genome. Biochim Biophys Acta Rev Cancer 1869(2):333–345

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13(5):278–289

Roberts RJ, Vincze T, Posfai J, Macelis D (2010) REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 38(Database issue):D234–D236

Roberts RJ, Vincze T, Posfai J, Macelis D (2015) REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 43(Database issue):D298–D299

Romano KA, Rey FE (2018) Is maternal microbial metabolism an early-life determinant of health? Lab Anim (NY) 47(9):239–243

Schadt EE, Banerjee O, Fang G, Feng Z, Wong WH, Zhang X, Kislyuk A, Clark TA, Luong K, Keren-Paz A, Chess A, Kumar V, Chen-Plotkin A, Sondheimer N, Korlach J, Kasarskis A (2013) Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res 23(1):129–141

Srikhanta YN, Maguire TL, Stacey KJ, Grimmond SM, Jennings MP (2005) The phasevarion: a genetic system controlling coordinated, random switching of expression of multiple genes. Proc Natl Acad Sci USA 102(15):5547–5551

Srikhanta YN, Gorrell RJ, Steen JA, Gawthorne JA, Kwok T, Grimmond SM, Robins-Browne RM, Jennings MP (2011) Phasevarion mediated epigenetic gene regulation in *Helicobacter pylori*. PLoS One 6(12):e27569

Takahashi K (2014) Influence of bacteria on epigenetic gene control. Cell Mol Life Sci 71 (6):1045–1054

Takahashi N, Naito Y, Handa N, Kobayashi I (2002) A DNA methyltransferase can protect the genome from postdisturbance attack by a restriction-modification gene complex. J Bacteriol 184 (22):6100–6108

Tavazoie S, Church GM (1998) Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in *E. coli*. Nat Biotechnol 16(6):566–571

Thanbichler M, Wang SC, Shapiro L (2005) The bacterial nucleoid: a highly organized and dynamic structure. J Cell Biochem 96(3):506–521

Tong T, Chen S, Wang L, Tang Y, Ryu JY, Jiang S, Wu X, Chen C, Luo J, Deng Z, Li Z, Lee SY, Chen S (2018) Occurrence, evolution, and functions of DNA phosphorothioate epigenetics in bacteria. Proc Natl Acad Sci USA 115(13):E2988–E2996

Tretyakova N, Villalta PW, Kotapati S (2013) Mass spectrometry of structurally modified DNA. Chem Rev 113(4):2395–2436

Vasu K, Nagaraja V (2013) Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol Mol Biol Rev 77(1):53–72

Wang L, Chen S, Vergin KL, Giovannoni SJ, Chan SW, DeMott MS, Taghizadeh K, Cordero OX, Cutler M, Timberlake S, Alm EJ, Polz MF, Pinhassi J, Deng Z, Dedon PC (2011) DNA phosphorothioation is widespread and quantized in bacterial genomes. Proc Natl Acad Sci USA 108(7):2963–2968

Wang L, Jiang S, Deng Z, Dedon PC, Chen S (2019) DNA phosphorothioate modification-a new multi-functional epigenetic system in bacteria. FEMS Microbiol Rev 43(2):109–122

Weis AM, Storey DB, Taff CC, Townsend AK, Huang BC, Kong NT, Clothier KA, Spinner A, Byrne BA, Weimer BC (2016) Genomic comparisons and zoonotic potential of campylobacter between birds, primates, and livestock. Appl Environ Microbiol 82:7165–7175

Westra ER, Swarts DC, Staals RH, Jore MM, Brouns SJ, van der Oost J (2012) The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. Annu Rev Genet 46:311–339

Wilson GG (1991) Organization of restriction-modification systems. Nucleic Acids Res 19 (10):2539–2566

Wion D, Casadesus J (2006) N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. Nat Rev Microbiol 4(3):183–192

Yang MK, Ser SC, Lee CH (1989) Involvement of *E. coli* dcm methylase in Tn3 transposition. Proc Natl Sci Counc Repub China B 13(4):276–283

Yotani T, Yamada Y, Arai E, Tian Y, Gotoh M, Komiyama M, Fujimoto H, Sakamoto M, Kanai Y (2018) Novel method for DNA methylation analysis using high-performance liquid chromatography and its clinical application. Cancer Sci 109(5):1690–1700

Eukaryotic Pangenomes

Guy-Franck Richard

Abstract The first eukaryotes emerged from their prokaryotic ancestors more than 1.5 billion years ago and rapidly spread over the planet, first in the ocean, later on as land animals, plants, and fungi. Taking advantage of an expanding genome complexity and flexibility, they invaded almost all known ecological niches, adapting their body plan, physiology, and metabolism to new environments. This increase in genome complexity came along with an increase in gene repertoire, mainly from molecular reassortment of existing protein domains, but sometimes from the capture of a piece of viral genome or of a transposon sequence. With increasing sequencing and computing powers, it has become possible to undertake deciphering eukaryotic genome contents to an unprecedented scale, collecting all genes belonging to a given species, aiming at compiling all essential and dispensable genes making eukaryotic life possible.

In this chapter, eukaryotic core- and pangenomes concepts will be described, as well as notions of closed or open genomes. Among all eukaryotes presently sequenced, ascomycetous yeasts are arguably the most well-described clade and the pangenome of *Saccharomyces cerevisiae*, *Candida glabrata*, *Candida albicans* as well as *Schizosaccharomyces* species will be reviewed. For scientific and economical reasons, many plant genomes have been sequenced too and the gene content of soybean, cabbage, poplar, thale cress, rice, maize, and barley will be outlined. Planktonic life forms, such as *Emiliana huxleyi*, a chromalveolate or *Micromonas pusilla*, a green alga, will be detailed and their pangenomes pictured. Mechanisms generating genetic diversity, such as interspecific hybridization, whole-genome duplications, segmental duplications, horizontal gene transfer, and single-gene duplication will be depicted and exemplified. Finally, computing approaches used to calculate core- and pangenome contents will be briefly described, as well as possible future directions in eukaryotic comparative genomics.

G.-F. Richard (✉)
Institut Pasteur, Department Genomes & Genetics, Paris, France

CNRS, Paris, France
e-mail: gfrichar@pasteur.fr

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_12

Keywords Eukaryotes · Evolution · Fungi · Haptophyte · Mamiellale · Eudicotyledon · Monocotyledon · Protostomian · Metazoan

1 The Origin of Eukaryotes

Respiratory-competent eukaryotic cells emerged more than 1.5 billion years ago, from the endosymbiosis of an alphaproteobacterium and an ancestral archaebacterium, probably belonging to the Asgard clade (Zaremba-Niedzwiedzka et al. 2017). This protoeukaryote evolved, concomitantly, a complex system of membrane compartments that would ultimately lead to the isolation of the genomic content within a real nucleus (*eu karyon* in Greek) while the degenerated alphaproteobacteria gave rise to the mitochondria (López-García and Moreira 2006). The subsequent acquisition of photosynthesis through endosymbiosis with a cyanobacteria evolved this primitive cell into a protoalga from which all plants will eventually develop. The general outline of this scenario has been postulated for more than a century (Mereschowsky 1999; Sagan 1967) and modern-day DNA sequencing techniques allowed to precisely identify bacteria most closely related to modern eucaryotes, hence representing their most probable ancestors. However, the exact order of events is still a matter of debate among evolution specialists. Did membranes come first, to isolate nucleic acid metabolism from protein and sugar metabolism? Did the mitochondria come first, providing a considerable source of oxidative energy to further develop a complex network of membranes? These two scenarios are not necessarily exclusive and one may also imagine that a number of different protoeucaryotes emerged at roughly the same time (at geological scale) and competed with each other within similar ecological niches, until one lineage arose and was eventually selected to give rise to all eukaryotic life.

Given the bacterial origin of nucleated cells, it was assumed that most if not all eukaryotic gene families would share homology to prokaryotic genes. However, the sequencing of an old deep-branching eukaryote, the excavata *Naegleria gruberi* (Fig. 1), revealed that only 57% of its 4133 protein families had a clear prokaryotic homologue. The remaining genes showed no homology to bacterial sequences and therefore appear to be eukaryote inventions. Therefore, one must expect eukaryotic pangenomes to be significantly different from any known prokaryotic pangenome.

2 Sequencing Eukaryotic Genomes

Modern-day eukaryotes are estimated to represent 8,740,000 land species and 2,210,000 ocean species, for a total of roughly 11 million, one order of magnitude above procaryotes (Mora et al. 2011). Higher estimates, based on plankton sampling, suggest figures around 16 million of oceanic eukaryotes and 60 million of land species (de Vargas et al. 2015). Eukaryote classification is a complex problem taking

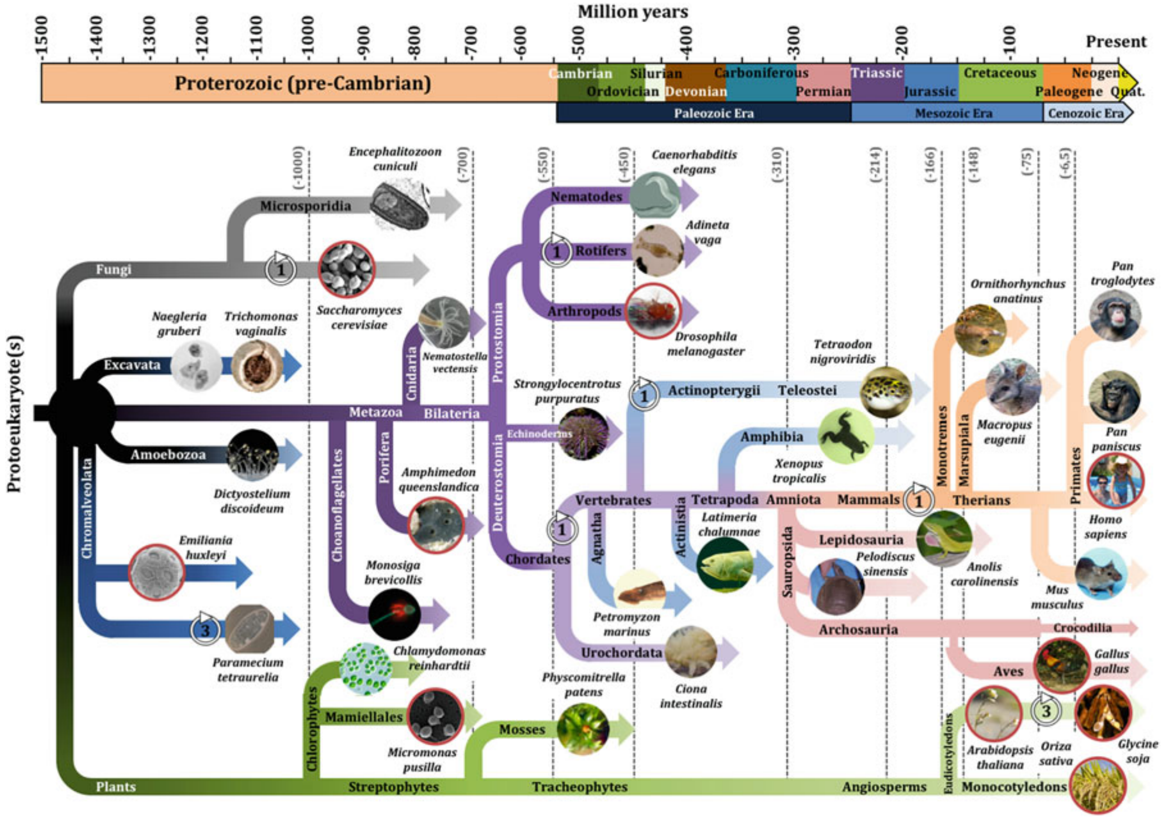


Fig. 1 Overview of some of the most representative eukaryotic genomes sequenced. At the top is shown a timeline in million years before present (note that the scale is different before and after Cambrian). Geological periods are indicated, as well as their corresponding eras. Each arrow represents one monophyletic

its roots into the nineteenth century zoology and botanics, but more recently gained much insight from whole-genome sequencing and molecular phylogeny reconstruction methods (Felsenstein 2004). Early eukaryotes (or old eukaryotes), such as fungi, monocellular green algae, excavata (one of the most basal lineage), amoebozoa, and chromalveolata diverged probably between 1.2 and 1.45 billion years ago (Embley and Martin 2006). Younger eukaryotes, like vertebrates, emerged 450 million years ago (Erwin et al. 2011), whereas *Homo sapiens* is still in evolutionary infancy with an estimated date of divergence from chimpanzee around 6.5 million years ago (Green et al. 2010) (Fig. 1).

The ascomycete *Saccharomyces cerevisiae* was the first eukaryote whose nuclear genome was totally sequenced, more than 20 years ago (Goffeau et al. 1996). In the 1990s, it took the efforts of 633 scientists from more than 100 laboratories during 8 years to complete it (Goffeau et al. 1997). In the modern genomic era, sequencing is fast, cheap, and allows to decipher whole eukaryotic genomes at unprecedented scale and pace in human history. At the present time, 707 different eukaryote species, including 54 unicellular animals (Protozoa) or algae, 300 metazoans (multicellular animals), 137 plants, and 216 fungi had their genome sequenced to various levels of completion and assembly. Indeed, the actual pace at which eukaryotes are being sequenced is so elevated, that the aforementioned figures will be completely outdated when this book will be published. Remarkably, one of the most ambitious current genome projects envisions to sequence all eukaryotic life present on planet Earth, and the cost of such a project would be similar to what was spent to sequence the first human genome alone (Pennisi 2017). Some of the most representative eukaryote species, whose genomes were completely sequenced are represented in Fig. 1, on the evolutionary branch they belong to, along with their estimated geological period of appearance based on molecular clocks.

Fig. 1 (continued) group (or clade) that survived to present day. Branch lengths are arbitrary. When more than one organism was sequenced in a given clade, only one was shown (for example, among all sequenced bird genomes only the paradigmatic *Gallus gallus* species was represented). Vertical dotted lines indicate speciation time from the most recent common ancestor, calculated from molecular clocks. For example, Actinopterygians (bony fish) separated from other vertebrates approximately 450 million years ago. Note that Precambrian radiation datations were only tentatively attributed, given the large uncertainties associated to ancient eukaryotes. Circled numbers represent whole-genome duplications detected by sequencing. The constriction between Archosaurians and Aves represents the Archaeopteryx, the ancestor of all modern birds (Hillier et al. 2004). The smaller arrow between Archosauria and Crocodilia represents the dinosaurian mass extinction, 66 million years ago, among whom the only survivors were the ancestors of modern-day crocodiles (Brugger et al. 2017; Renne et al. 2015). Red circled species were used to define core- and pangenomes and are more extensively described in the text

3 The 1000 Genome Projects

One of the most remarkable aspects of modern-day genomics is the ambition to describe a large number of individuals (usually in the range of thousands) belonging to the same monophyletic group (or clade). When the first eukaryotic genome sequences were completed, it became apparent that one genome would not be sufficient to describe the whole species. Several programs subsequently started, aiming at sequencing a large number of individuals belonging to the same species and comparing them to the first genome, usually called "reference genome" because its state of completion and annotation was often more advanced. Several of these projects have been completed over the last few years: 1011 *S. cerevisiae* genomes (Peter et al. 2018), 1135 *Arabidopsis thaliana* genomes (The 1001 Genomes Consortium 2016), 2504 followed by 10,545 human genomes (Telenti et al. 2016; The 1000 Genomes Project Consortium 2015), and 1483 rice genomes (Yao et al. 2015) have already been sequenced, but complete analyses of gene content and core- and pangenome calculations are not always published. Even more ambitious endeavors are planned: the 10,000 plant genome project led by the Chinese BGI[1] aims at sequencing one representative plant from every major clade (Normile et al. 2017); the same institute launched in 2015 the 10,000 bird genome project, in an attempt to sequence every one of the 10,500 living bird species (Zhang 2015). The i5K initiative is planning to sequence 5000 arthropod genomes (i5K Consortium 2013) or the Genome 10K project intends to sequence 10,000 vertebrate genomes (Genome 10K Community of Scientists 2009). All these projects—and many others to come—will contribute to unraveling the complete set of genes used by eukaryotic life forms on Earth. With this wealth of data at hand, assuming it will not be too overwhelming for available data storage and computing power, essential questions should find their answers. What are the core genes shared by all eukaryotic species? How many different versions of the same gene (alleles) can be found? How many variable or dispensable genes can be detected in a given species? What is the size of a species pangenome, of a clade pangenome, of the eukaryotic pangenome itself?

4 Defining Eukaryotic Pangenomes: Open or Closed?

The very notion of pangenome was coined by Hervé Tettelin and colleagues in a 2005 seminal article, describing sequencing and genome analysis of eight strains of *Streptococcus agalactiae*. Despite a high degree of synteny[2] between isolates, the authors detected 69 genomic islands that were absent in at least one genome, some characterized by an atypical nucleotide compositional bias, suggestive of a possible acquisition by horizontal transfer. They showed that the number of shared genes in all

[1] Beijing Genomics Institute, the largest—by far—sequencing center in the world.

[2] Synteny: gene order along a chromosome.

species decreased at each addition of a new genome, reaching the minimal number of 1806 genes. On the contrary, each genome addition increased the number of variable genes, those that are absent in one or more strain. They proposed that a bacterial species may be defined by a set of genes present in all strains (core-genome) and by a dispensable—or variable—set of genes, composed of those present in at least one strain but absent from all others. The addition of these variable genes to the core-genome would make what was called the "pangenome" (from the Greek word *pan (παν)*, meaning "whole") (Tettelin et al. 2005). Mathematical modeling showed that the pangenome measurement followed the Heap's law, an empirical law used in information retrieval, in which as more and more books are read, the number of different words grows as a power law of the total number of books read. The function form of the power law depends on two parameters: the exponent α and a proportionality constant. Practically, the number of new genes discovered after each new genome sequence will be: $n = \kappa N^{-\alpha}$, in which κ is a constant, N is the number of genomes sequenced, and $\alpha > 0$. For $\alpha > 1$, the pangenome size approaches a plateau as more and more genomes are sequenced, the pangenome is "closed" (Fig. 2a). On the other hand, for $0 < \alpha \leq 1$, the pangenome size will increase at each new genome addition and the pangenome is "open" (Fig. 2b) (Tettelin et al. 2008).

Among sequenced bacterial species, some exhibit a closed pangenome, for example *Staphylococcus aureus* ($\alpha = 1.84$), *Streptococcus pyogenes* ($\alpha = 1.88$), *Ureaplasma urealyticum* ($\alpha = 2.5$) or the extreme case of *Bacillus anthracis* ($\alpha = 5.6$). Others display an open pangenome, like *Bacillus cereus* ($\alpha = 0.65$) or the cyanobacteria *Prochlorococcus marinus* ($\alpha = 0.80$). Note than when α is equal or very close to 1, the pangenome is still open, but the rate of acquisition of new genes is very slow. This is the case of *Escherichia coli* ($\alpha = 1.04$), *Streptococcus agalactiae* ($\alpha = 1.05$), or *Streptococcus pneumoniae* ($\alpha = 0.98$) (Tettelin et al. 2008).

5 Yeast Pangenomes

5.1 Saccharomyces cerevisiae

Historically, budding yeast was the first eukaryote whose genome was completely sequenced (Goffeau et al. 1996). A British collaborative work in which 70 *S. cerevisiae* and *S. paradoxus* isolates were sequenced to low coverage showed that *S. cerevisiae* strains showed less variability than *S. paradoxus* strains. Worldwide budding yeast population structure was made of a few geographically isolated lineages and of several mosaic genomes, and underlined the possibility that humans played a major role in producing these variations by transporting and selecting yeast strains (Liti et al. 2009). Following this pioneering work, a collaborative effort of two French laboratories and the Genoscope led to the completion of 1011 *S. cerevisiae* isolates, collected worldwide, from domesticated, wild, or human origin (mainly clinical). This sequencing effort allowed to determine that Chinese and Taiwanese

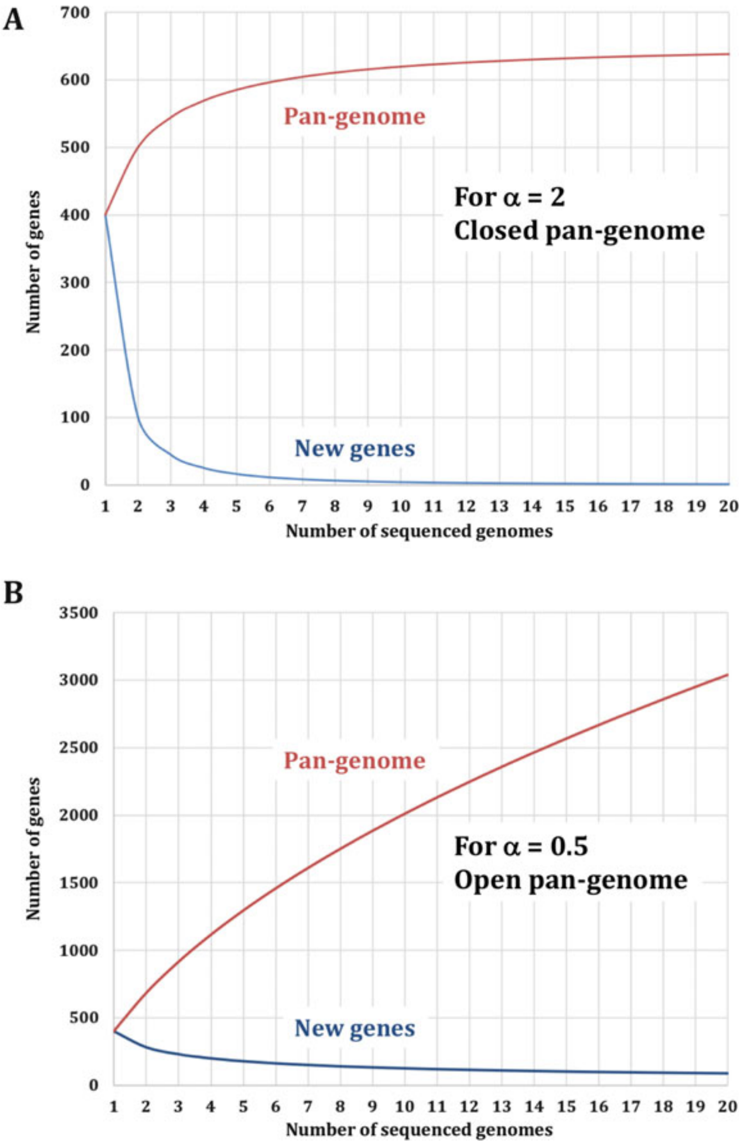


Fig. 2 Open versus closed pangenomes. (**a**) Closed pangenome. In this example, the number of new genes = 400× (Nbr genomes)$^{-\alpha}$, with $\alpha = 2$. The number of new genes revealed by each new genome sequence rapidly decreases and the pangenome size reaches a plateau. (**b**) Open pangenome. The number of new genes = 400× (Nbr genomes)$^{-\alpha}$, with $\alpha = 0.5$. The number of new genes revealed by each new genome sequence keeps on growing and the pangenome size steadily increases

strains were closer to *Saccharomyces paradoxus* and to the root of the *Saccharomyces sensu stricto* than strains from any other origin, strongly supporting a single out-of-China origin for *S. cerevisiae*, that subsequently spread all over the planet. Using de novo assembly and a specific detection pipeline, it could be determined that the yeast core-genome contained 4940 Open Reading Frames (ORFs) whereas 2856 ORFs were variable within the population, for a total of 7796 ORFs constituting the pangenome (Peter et al. 2018) (Table 1). Core ORFs were mostly found in one copy per haploid genome, while ca. 20% of variable ORFs were absent or present in more than one copy. The authors subsequently looked at the origin of these variable ORFs and classified them in three different groups, based on their phylogeny: ORFs with their closest ortholog in another *S. cerevisiae* strain and consistent with genome phylogeny were considered as being ancestral acquisitions; ORFs with their best ortholog in another *Saccharomyces* species were considered to be introgressions; and finally ORFs more related to another yeast species outside the *Saccharomyces* complex were treated as horizontal gene transfers (HGT) (Fig. 3a). Using these definitions, 1380 variable ORFs were assigned to an ancestral inheritance, 913 were designated as introgressions, and 183 were likely to be the result of HGT events from distant relative yeast species. Half of these HGT ORFs could be traced to *Torulaspora* or *Zygosaccharomyces* species. Given that these yeasts share similar environmental fermentative niches, it is likely that such physical promiscuity favored frequent transfer of genetic material between these species. In six cases, large HGT events (38–165 kb) were identified, but most isolates retained only mosaics of small segments suggesting that the large ancestral HGT underwent several rounds of successive deletions leading to the complex patterns observed today. Among the 913 introgressions, 97% were unambiguously acquired from *S. paradoxus*, all *S. cerevisiae* ORF carrying at least one *S. paradoxus* ORF, suggesting continuous gene flows between these two yeast species. This is in good agreement with a former work using microarrays to genotype *Saccharomyces* strains of different origins, in which most introgressions detected in *S. cerevisiae* came from *S. paradoxus* (Dunn et al. 2012). Finally, two-thirds of ancestral acquisitions were present in at least half the yeast isolates, suggesting that they segregated in most strains since the time of their acquisition (Fig. 3b).

The core- and pangenomes of the S288C reference strain were analyzed more thoroughly for variable gene functions. Out of 6081 ORFs, 1144 were identified as variable. The distribution of these ORFs was found to be skewed toward subtelomeric regions, which have been known for a long time to be highly polymorphic among yeast strains and species (Fabre et al. 2005). Functions of variable ORFs were strongly enriched for cell-wall and membrane components, cell–cell interactions, and secondary metabolism. Finally, core-genome ORFs were found to exhibit lower levels of loss-of-function mutations, as compared to pangenome ORFs, as well as a lower dN/dS ratio of nonsynonymous over synonymous substitutions, showing that the former were less constrained than the latter.

Table 1 Core- and pangenome contents

Clade	Species (or genus)	Isolates	Core-genome	Variable genes[a]	Pangenome	Status[b]
Saccharomycotina	*Saccharomyces cerevisiae*	1011	4940	2856 (37%)	7796	ND
	Candida glabrata	33	3603	9915 (73%)	13,000–14,000	ND
	Candida albicans	21	6069	120 (2%)[c]	6189	ND
Taphrinomycotina	*Schizosaccharomyces*[d]	4	4218	782 (16%)	5000	ND
Eudicotyledon	*Glycine soja*	7	28,716	30,364 (61%)	50,080	Open
	Brassica oleracea	9	49,895	11,484 (19%)	61,379	ND
	Populus trichocarpa	6	≈34,000	12,000–13,000 (26%)	46,000–47,000	Closed
	Arabidopsis thaliana	19[e]	26,373	11,416 (30%)	37,789	Open
Monocotyledon	*Oriza sativa*	66	26,372	16,208 (38%)	42,580	Closed
	Zea mays	503	16,393	25,510 (61%)	41,903	Closed
	Hordeum vulgare	16	10,922	17,840 (62%)	28,762	Closed
Mamiellales	Three different species[f]	4	7137	2824 (23%)	12,518	ND
Haptophyte	*Emiliania Huxleyi*	14	20,055	10,514 (34%)	30,569	ND
Protostomian	*Drosophila*[g]	12	6698	40,852 (86%)	47,550	ND
Metazoans	*Homo sapiens*	5[h]	ND	ND	ND	Open

ND: Not Determined by the authors and not possible to calculate from published data

[a] The proportion of variable genes as compared to the pangenome size is indicated in parenthesis

[b] Open or closed pangenome (see Fig. 2 and text)

[c] Calculated from an average value. The real number of variable genes might be slightly larger

[d] *S. octosporus*, *S. pombe*, *S. japonicus* and *S. cryophilus*

[e] More than 1100 *A. thaliana* genomes were sequenced, but 19 transcriptomes were used to determine core-and pangenome contents (see text)

[f] Two isolates of *Micromonas pusilla*, plus *Ostreococcus tauri* and *Ostreococcus lucimarinus*

[g] The 12 sequenced genomes corresponded to 12 *Drosophila* species, not 12 isolates from the same species

[h] More than 10,000 human genomes are available, but 5 of them serve as references (see text)

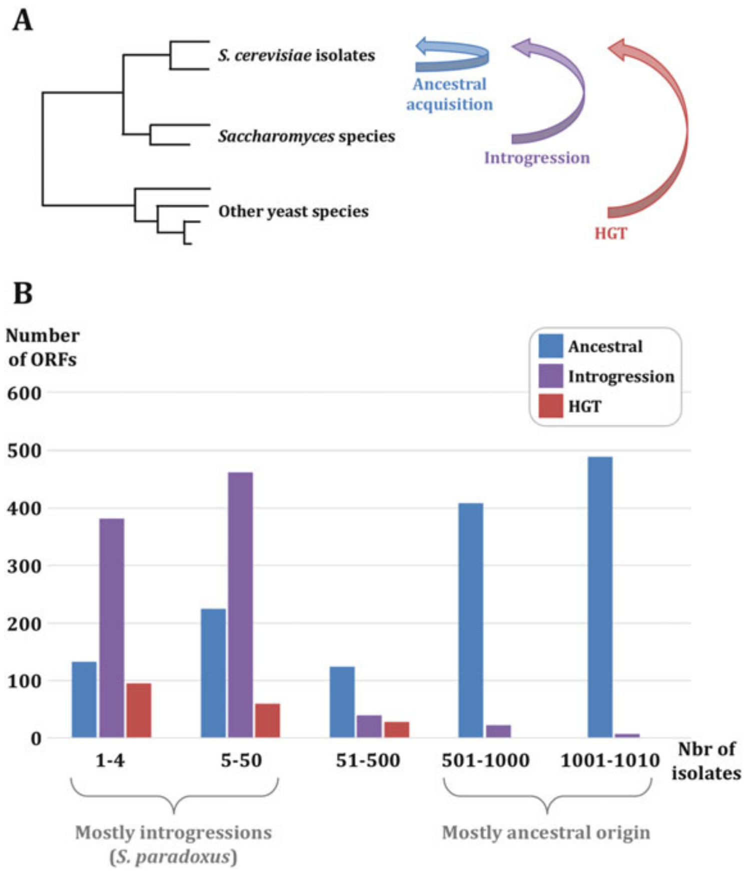


Fig. 3 Variable ORFs of the *S. cerevisiae* pangenome. (**a**) Phylogenetic origin of variable ORFs. ORFs were considered ancestral acquisitions when the best match was found to be a *S. cerevisiae* ORF (blue arrow), it was treated as an introgression when the best homolog was another *Saccharomyces* species (purple arrow), or a horizontal gene transfer (HGT, red arrow) when it was found to be another yeast species. (**b**) Distribution of variable ORFs. The number of isolates is indicated on the *X*-axis and the number of variable ORFs in each category is represented on the *Y*-axis

5.2 Candida glabrata

C. glabrata is an opportunistic pathogen responsible for candidiasis and bloodstream infections in immunocompromised patients (Bodey et al. 2002). It is the second cause of nocosomial infections, after *Candida albicans*, and a growing concern in public health, due to its resistance to azole antifungal drugs (Pfaller and Diekema 2004). Despite its genus name, its genome is closer to *S. cerevisiae* than to *C. albicans*. It belongs to the *Nakaseomyces* clade that also includes *Candida nivariensis* and

Candida bracarensis, two emerging pathogens, as well as *Nakaseomyces delphensis*, *Nakaseomyces bacilisporus*, and *Candida castellii*, three nonpathogenic species (Fig. 4). Comparison of orthologous proteins conservation shows that this clade is as distant from the *Saccharomyces* clade as man is distant from fish (Dujon 2006). Hence, the distance between orthologous proteins belonging to these two monophyletic groups is similar to the distance covered by vertebrate proteins since the actinopterygian radiation, some 450 million years ago[3] (Fig. 1). *C. glabrata* exhibits frequent chromosome polymorphisms among different isolates, due to translocations, copy number variations (CNV), gene tandem amplifications (Muller et al. 2009), formation of neo-chromosomes (Polakova et al. 2009), and the presence of many large tandem repeats known as megasatellites (Rolland et al. 2010; Thierry et al. 2008, 2009). The five aforementioned pathogenic and nonpathogenic *Nakaseomyces* species were sequenced to high coverage and their sequence was compared to the *C. glabrata* CBS138 reference strain (Dujon et al. 2004). Protein contents range from 4875 for *C. castellii* to 5315 for *C. bracarensis*, figures significantly lower than the 5886 *S. cerevisiae* proteins (Gabaldon et al. 2013). Among gene losses in *Nakaseomyces*, four entire multigene families (*PHO*, *SNZ*, *SNO*, and *PAU*) were absent in all species or represented by only one member in *C. castellii* or *N. bacillisporus*. These genes are involved in phosphate metabolism (PHO), in nutrient limitation response (SNZ and SNO), or in alcoholic fermentation (PAU). The loss of BNA genes, functioning in de novo synthesis of nicotinic acid probably results from the yeast adaptation to its human host, since colonization of the urinary tract occurs through induction of adhesin genes, upregulated in nicotinic acid-poor medium, such as urine (Domergue et al. 2005). The *C. glabrata* genome contains a large number of genes that are absent from *S. cerevisiae* and specifically involved in adhesion and virulence. The *EPA* genes, a family of glycosyl-phosphatidylinositol cell-wall genes, completely absent from *S. cerevisiae*, was represented by 18 members in the *C. glabrata* reference strain (CBS138), and seven additional genes were present in the BG2 strain, widely used in adhesion studies (Cormack et al. 1999). Remarkably, the two other pathogenic species, *C. bracarensis* and *C. nivariensis*, contained respectively 12 and 9 members of the EPA family, whereas the nonpathogenic *N. delphensis* and *C. castelli* harbored respectively one and three copies and *N. bacillisporus* presented only one distant homologue. In addition, the *C. glabrata* genome contained 44 genes comprising internal repeats, whose motifs were 135–300 nt long, tandemly repeated 3–30 times in frame (Thierry et al. 2008). These megasatellites encode many serine and threonine residues and genes harboring these tandem repeats were proposed to encode cell-wall glycoproteins and to be involved in cellular adhesion (Thierry et al. 2009). Phylogenetic studies of 21 fungal genomes showed that these megasatellites were uniquely found in *C. glabrata*, but their presence among other members of the *Nakaseomyces* has not been tested yet (Tekaia et al. 2013).

[3] This does not mean that *Saccharomyces* and *Nakaseomyces* diverged 450 million years ago, because there is no reliable molecular clock for yeasts.

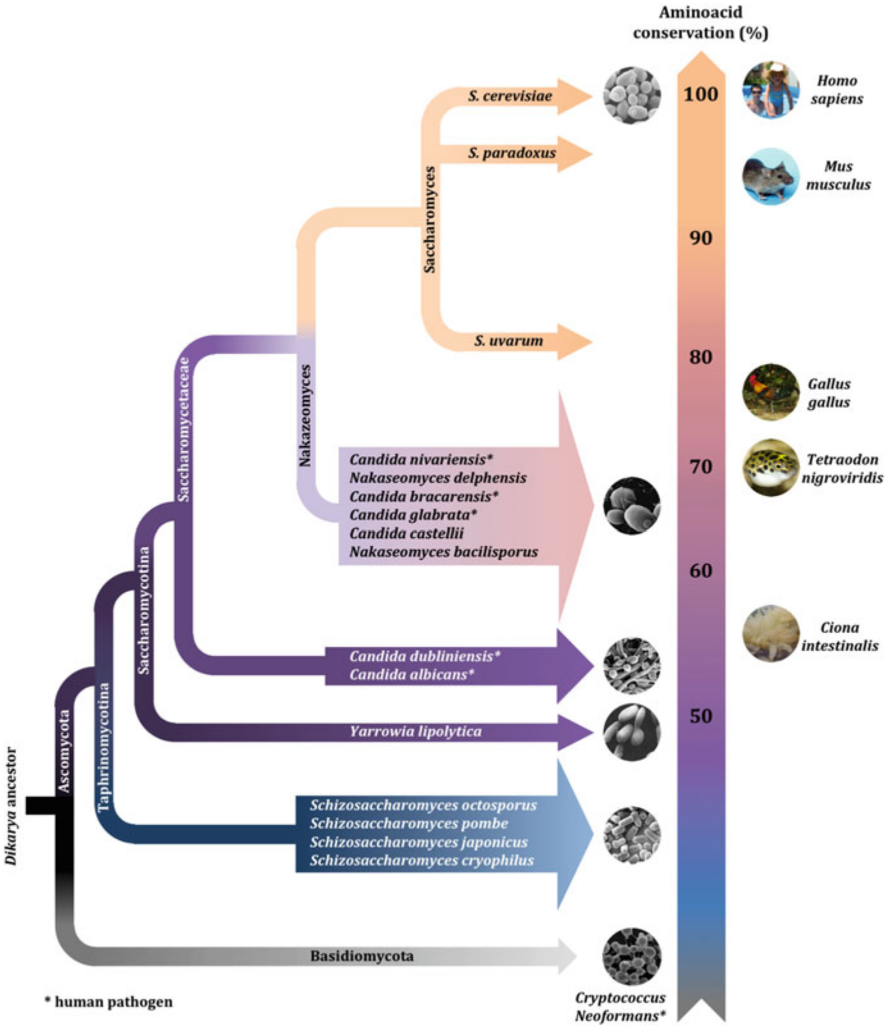


Fig. 4 Yeast pangenomes of the *Dikarya* tree. On the left, the figure shows some of the yeast species whose genomes were completely sequenced, arranged by clade. Branch lengths are arbitrary and do not reflect evolutive distances. On the right, amino acid conservation of orthologous proteins between yeast and between animal species are indicated (adapted from Dujon 2006)

In a very recent study, 33 isolates of *C. glabrata* of different geographical origins were fully sequenced and compared to the CBS138 reference strain (Carreté et al. 2018). Altogether, 108 genes were deleted or duplicated in these strains, half of them encoding glycosylphosphatidylinositol-anchored adhesin homologues, showing the extensive variability of this gene family within this clade. The core-genome contained 3603 proteins, significantly less than for *S. cerevisiae* (see above). On the contrary, the number of variable ORFs was higher than budding yeast, since

302–580 predicted genes (mean: 342) were found to be unique of each isolate, for a total of 9915 strain-specific genes among 29 strains considered.[4] This figure may be partially overestimated, due to automated annotations or clustering artifacts, but from these data one may infer that the *C. glabrata* pangenome covers 13,000–14,000 genes, almost twice as many as the *S. cerevisiae* pangenome.

In conclusion, yeasts of the *C. glabrata* clade contain significantly fewer genes than *S. cerevisiae*, with specific gains and losses as compared to their distant cousin. However, gene content is highly variable among *Nakaseomyces* and the *C. glabrata* pangenome size is larger than the *S. cerevisiae* pangenome, although further analyses are needed to narrow down these numbers.

5.3 Schizosaccharomyces *Genomes*

Fission yeasts are very distant relatives of *S. cerevisiae* and the *Taphrinomycotina* clade comprise only four known species: *Schizosaccharomyces japonicus*, *Schizosaccharomyces cryophilus*, *Schizosaccharomyces octosporus*, and the model yeast *Schizosaccharomyces pombe*. They form a basal branch of the *Dikarya*[5] tree (Fig. 4) and exhibit very distinct life history and metabolism as compared to *Saccharomycotina*. Under many aspects, *S. pombe* is actually closer to metazoans than to budding yeasts: among the more prominent features, large repetitive centromeres, heterochromatin histone methylation, heterochromatin proteins, RNA interference, telomere-binding proteins, cell-cycle control, the mitochondrial translation code, splicing and spliceosome components are more similar to metazoans. In addition, core orthologous genes in *S. pombe* are closer to metazoan genes than to other *Ascomycota*. Phylogeny reconstruction of the clade using high coverage sequence of the four *Schizosaccharomyces* species and 440 single-copy core orthologues surprisingly revealed that *S. pombe* and *S. japonicus* were as far to each other (55% average amino acid identity) as man and *Ciona intestinalis*, an urochordate (Fig. 1) (Rhind et al. 2011). The two other species, *S. octosporus* and *S. cryophilus*, were closer to each other (85% amino acid identity). Retrotransposons are numerous in *S. japonicus* and sequence divergence of their reverse transcriptase suggests that they predate the last ancestor of the *Ascomycota*. However, transposons were dramatically lost in the three other species, since *S. pombe* harbors two related retrotransposons, *S. cryophilus* contains only one and *S. octosporus* only has sequence relics of reverse transcriptase sequences. This loss was accompanied by a reorganization of centromere architecture, replacing the numerous transposons found at *S. japonicus* centromeres by other kinds of repeated sequences unrelated to transposons and specific of each of the other three species.

[4] Four isolates were excluded from this analysis because of low-quality assembly.

[5] *Ascomycota* and *Basidiomycota* together form the *Dikarya*.

Out of ≈5000 coding genes in fission yeasts, 4218 (84%) were identified as single-copy orthologues common to all four species. For some gene families, the level of conservation was even higher: 93% of protein kinases were common and more surprisingly 81% of introns (2901 out of 3601) were identical across the clade. Most gene gains were species- or clade-specific genes not found in another yeast species, whereas gene loss included the glyoxylate cycle, glycogen biosynthesis, the phosphoenolpyruvate carboxykinase, fewer *ADH* genes and lack of transcriptional regulators of glucose repression, all these changes reflecting the inability of fission yeast to use ethanol as a carbon source, although it produces it by fermentation. Hence, despite large evolutionary distances of conserved orthologous proteins, *Schizosaccharomyces* show a remarkably stable gene content, supporting a pangenome size only 10–20% larger than its core-genome.

5.4 Candida albicans

Candida albicans is another opportunistic pathogen, responsible for mucosal and systemic infections in immunocompromised patients. It is also a commensal of the gastrointestinal tract. Natural isolates of *C. albicans* are diploid and under specific conditions they are able to mate, resulting in tetraploid cells subsequently shifting to diploidy via random chromosome loss (Bennett and Johnson 2003). The nuclear genome of SC5314, a standard laboratory strain widely used in molecular analyses, was published in 2004. It revealed a high level of single-nucleotide polymorphisms (SNP) between both homologues, representing 90% of all detected polymorphisms, with an average frequency of one SNP in 237 bases. Heterozygosity was not homogeneous, since several chromosomes were interrupted by large regions of homozygosity (Jones et al. 2004). After that initial study, 21 clinical isolates of *C. albicans*, characterized by different phenotypic profiles, were also completely sequenced. Single-nucleotide polymorphisms were very limited among the isolates, being one order of magnitude lower than what was commonly found among *C. glabrata* strains (Gabaldón and Fairhead 2019). The gene content of these isolates was very similar to that of SC5314 reference strain, since most of its genes were present in all isolates (6069 genes out of 6189—or 98%—on the average), with few variable genes (Table 1). Genes exhibiting the most variable number of copies were retroelements as well as the subtelomeric *TLO* gene family. The position and number of *TLO* genes varied from 10 to 15 among isolates, indicative of a high level of plasticity (Hirakawa et al. 2015). More recently, the *Candida dubliniensis* genome, another opportunistic pathogen, less virulent than *C. albicans*, was sequenced. Except for translocations and chromosomal rearrangements that may be expected between two yeast species, both gene contents were found to be surprisingly similar. Out of 5569 orthologues, 5363 (96.3%) were more than 80% identical at the nucleotide level, and synteny was conserved for 98% of genes (Jackson et al. 2009). The search for species-specific genes identified 111 ORFs in *C. dubliniensis* and 191 in *C. albicans*. However, most of these variable ORFs corresponded to transposable elements. When these were filtered off, the real number of species-specific genes dropped to 29 and 168, respectively. Among those, the *TLO* gene family (12 members in *C. albicans*) was specifically expanded in this species, since only two copies were detected in *C. dubliniensis* and species-specific copies were monophyletic, supporting an independent expansion in *C. albicans*. On the contrary, the *IFA* gene family (13 members in *C. albicans*) underwent massive gene loss in *C. dubliniensis*, since several gene relics at various stages of decay were identified in this yeast species. In conclusion, in the present state of analysis, it appears that the core-genome common to *C. albicans* and *C. dubliniensis* probably approximates 5400 genes and that their pangenome may be predicted to be slightly larger, possibly around 6200 genes.

6 Plant Pangenomes

6.1 Soybean Genomes

Glycine max is the cultivated soybean variety, whose genome was published in 2010 (Schmutz et al. 2010). It was domesticated 5500 years ago and has been under intensive selection by human populations for yield increase. It diverged from the wild variety, *Glycine soja*, 800,000 years ago, well before its domestication. Therefore, natural selection contributed to differentiation of the two subspecies well before human selection started. In order to estimate the genetic diversity between domesticated and wild soybean species, the genome of seven *Glycine soja* isolates from south-east Asia were sequenced and compared to each other and to *G. max* (Li et al. 2014). Gene number ranged from 54,256 to 57,631, depending on the isolate and hundreds of genes were identified as gained or lost as compared to domesticated soybean. The *G. soja* core-genome contained 28,716 genes, while 30,364 variable genes were identified. Most of them (58%) were shared by two to six out of seven samples, whereas 12,916 (42%) were uniquely found in one of the seven isolates. The pangenome therefore contained 50,080 genes and covered 986.3 Mb of sequence. Its size increased with each new isolate, but it did not reach an asymptote, suggesting that adding new isolates would increase pangenome size (Fig. 2). Interestingly, dispensable genes exhibited more sequence variability than core genes. SNP frequency was at 2.67 sites per kilobase for variable genes, whereas it was significantly higher for core genes (4.12 sites per kilobase), and a similar bias was found for indels. Biological processes enriched in dispensable genes include specific metabolic processes, antioxidant activity, and structural molecule activity. These genes were also less conserved than core genes since 58% could not be assigned to a functional annotation, as compared to only 34% of the core genes. Lineage-specific genes include 11 genes implicated in effector-triggered immunity, acting as pathogen detectors, reflecting adaptation to various biotic stresses.

The domesticated soybean genome contains 1794 genes involved in acyl lipid metabolism, illustrating the effect of its intense selection for oil and fatty acid production. Among those, 32 exhibited CNV when compared to *Glycine soja*, 252 contained SNPs or indels and 21 showed high dN/dS ratios, suggestive of their possible positive selection in *Glycine max*.

In conclusion, *G. soja* pangenome was found to be twice as large as its core-genome, and its comparison with the domesticated *G. max* species revealed the effect of human selection on this widely cultivated crop.

6.2 Rice Genomes

Rice (*Oryza sativa* L.) is one of the most important crops in the world, feeding half the world population. The genome sequence of this monocotyledon was published in 2005 (International Rice Genome Sequencing Project 2005), although draft sequences of each chromosome were released earlier. Domesticated rice comprises two subspecies: *indica* and *japonica*. The reference genome (Nipponbare) is a *japonica* subspecies and contains 37,544 protein-coding genes, among which 2859 (8%) seemed to be uniquely found in rice. In an effort to explore the genetic diversity of cultivated rice, 1483 sequences of both subspecies from 73 countries, sequenced at low coverage (1–3 X), were compared to the reference genome. Comparison of both subspecies sequences to the reference genome identified 8991 predicted genes for the dispensable *indica* genome and 6366 for the *japonica* genome. Among these, strong evidence of expression or high homology was found for 1120 genes of the *japonica* dispensable genome and 1913 genes of the *indica* dispensable genome. Out of these 1913 high confidence genes, 1189 (62%) contained a recognizable protein domain, for a total of 276 different protein domains altogether (Yao et al. 2015).

In a more recent study, 66 isolates of cultivated rice as well as wild rice (*Oriza rufipogon*)[6] were sequenced to high coverage and the corresponding genomes were de novo assembled and compared (Zhao et al. 2018). Chromosomal introgressions from *indica* were detected in ≈16% of tropical *japonica* genomes. Numerous insertions and deletions were identified within genes, since a total of 10,872 genes were at least partially absent from the reference genome, due to large indels. Protein-coding genes present in at least one isolate were annotated and all transposable elements were filtered out. A total of 26,372 genes were found to be common to more than 60 rice isolates and were therefore considered to constitute the rice core-genome. Variable genes, present in less than 60 genomes, were assigned to a dispensable set of 16,208 genes, so that the rice pangenome reached a total of 42,580 genes. A larger proportion of core proteins (78%) than of dispensable proteins (36%) matched to known domains, suggesting that some of these variable genes may be pseudogenes or artifacts. Among dispensable genes, abiotic and biotic response genes, controlling disease resistance in rice were found to be enriched. When coding genes were sequentially added from each genome, the number of

[6] 28 *Oriza sativa japonica*, 25 *Oriza sativa indica*, and 13 *Oriza rufipogon* isolates.

different genes reached a plateau, although more pronounced for gene families than for singletons. This strongly suggests that the rice pangenome is almost closed and that further sequencing of rice isolates will not prove to be very useful in identifying new dispensable genes (Table 1).

6.3 Maize Genomes

Transcriptome sequencing of polyadenylated mRNAs was used in a genome-wide study as a proxy to determine the complete set of protein-coding genes within 503 diverse maize inbred isolates of different origins (Hirsch et al. 2014). RNA-seq reads were mapped to the *Zea mays* reference genome and reads that did not match were used for identification of novel transcripts. To limit redundancy, only the longest transcript of each locus was taken into consideration for further analysis. A total of 8681 high confidence transcripts that were absent from the reference genome were categorized as dispensable genes. Among those, 50% matched with rice and sorghum proteins, ruling out that they could be artifacts or contaminants. Transcripts detected in all isolates, including the reference line, represented 16,393 genes and constituted the core-genome. Dispensable transcripts, that were identified in only a subset of isolates, represented 25,510 genes, for a pangenome of 41,903 genes, very close to the rice pangenome, although the proportion of variable genes was much higher in maize (61% vs. 38% for rice). Sequential addition of genes belonging to each isolate revealed that the number of different singletons and gene families reach a plateau (more pronounced for singletons), demonstrating that the maize pangenome was closed, or very close to completion (Table 1).

6.4 Cabbage Genomes

Brassica oleracea is a diploid eudicotyledon, comprising remarkably morphologically diverse crops, including cabbage, cauliflower, broccoli, Brussels sprout, kohlrabi, and kale. The *B. oleracea* pangenome was built by sequencing nine isolates (eight cultivated and one wild—*Brassica macrocarpa*) and anchoring them on one of the two reference genomes (Parkin et al. 2014). The assembled pangenome covers 587 Mb and represents 61,379 genes, after removal of transposable elements. The core-genome constitutes the majority of the pangenome, representing 49,895 genes (81%), whereas 11,484 genes (19%) are variable, 1322 (2%) being present in only one line. Dispensable genes were enriched for functions predicted to be involved in disease resistance, defense response, water homeostasis, amino acid phosphorylation, and signal transduction. Lineage-specific variable genes comprised biotic and abiotic stress response genes, similar to what was observed in rice and soybean. *B. oleracea* underwent a whole-genome triplication specific to this lineage, in which gene families involved in auxin function and in morphological variations were amplified, these last ones perhaps contributing to the wide morphology diversity observed in this species.

There are 14 variable genes predicted to regulate flowering time and maturity in *B. oleracea*, but all of them were absent from one of the two reference strains (TO1000), a rapid cycler. One of the flowering loci, *FLC* (Flowering Locus C), is an important regulator of vernalization and regulates flowering time variation by the number of gene copies. One *FLC* gene was present in *Arabidopsis thaliana*, whereas four paralogues were found in *B. oleracea*. All four were part of the core-genome and two additional homologues were detected: one was present in all lines except the TO1000 reference strain and the other was present only in *B. macrocarpa* and one isolate (Cauliflower1). Independent functional studies showed that disruption of this gene in cauliflower led to early flowering, strongly suggesting that its absence in TO1000 was responsible for the early flowering of this rapid cycler (Golicz et al. 2016).

Genetic signatures of the core-genome and of the variable genome are very different. Core genes are longer on the average and harbor more exons. They also have lower mean SNP density and the ratio of non-synonymous over synonymous substitutions was lower than for variable genes, suggesting that core genes were under a more selective purifying selection than variable genes. In conclusion, *B. oleracea* core and variable genes exhibit the same properties that were observed in other eucaryotic pangenomes.

6.5 Poplar Genomes

The genome of *Populus trichocarpa*, black cottonwood, was published in 2006. Out of its predicted 45,555 protein-coding genes, 40,307 (88%) had a homologue in *Arabidopsis thaliana*, while conversely 91% of *A. thaliana* predicted genes showed some similarity to a *P. trichocarpa* gene (Tuskan et al. 2006). More recently, six isolates of other poplar species, four *Populus negra* and two *Populus deltoides*, were sequenced to 26-45X coverage and compared to the *P. trichocarpa* reference genome. Genome comparisons identified 7889 deletions and 10,586 insertions in the two newly sequenced species, as compared to *P. trichocarpa*. However, a large majority of these were due to transposons and retrotransposable elements (62% of deletions and 84% of insertions), a feature shared by all plant pangenomes sequenced so far. Once transposon sequences were filtered out, 3230 genes exhibiting CNV signatures between at least two of the samples were detected. These CNVs were significantly more abundant within 3 Mb from telomeres and corresponded to gene additions or deletions in one or more sample. A total of 230 variable genes were detected among *P. nigra* samples, and of 174 dispensable genes between the two *P. deltoides* isolates. The reference *P. trichocarpa* genome showed 187 genic variations with *P. nigra* and 213 with *P. deltoides*. Among these dispensable genes, 70% belonged to a gene family, allowing to detect some over-represented gene functions. Remarkably, variable genes were preferentially involved in signal transduction, receptor activity, and disease resistance, similarly to what was observed for soybean, rice, and cabbage (Pinosio et al. 2016).

The authors of this study calculated that the poplar pangenome was approximately 500 Mb, 80% being shared by all the isolates and therefore constituting the core-genome. When *P. nigra* and *P. deltoides* genomes were compared to the reference *P. trichocarpa*, 2270 genes were absent from at least one sample and 2453 other genes were detected in a variable number of copies, for a total of 4723 variable genes. Unfortunately, the proportion of dispensable genes between *P. nigra* and *P. deltoides* was not determined, and it was therefore not possible to figure out the exact size of the poplar pangenome. However, estimates suggest a size of ≈34,000 genes for the core-genome and ≈12,000–13,000 variable genes, giving a pangenome size of ≈46,000–47,000 genes. Using available data about *P. nigra* dispensable genes, it is tempting to suggest that its pangenome should be closed.

6.6 Mamiellales Genomes

Micromonas pusilla is a marine picoeukaryote of the Mamiellales order, measuring less than 2 μm and living in all oceans worldwide. Two independent isolates of *M. pusilla* were sequenced and their genomes were compared to those of *Ostreococcus lucimarinus* and *Ostreococcus tauri*, two other Mamiellales. Surprisingly, the two *Micromonas* shared only 90% of their 10,000 predicted genes, whereas the two *Ostreococcus* shared 97% of theirs. Comparison of the four sequences allowed to define a core-genome containing 7137 genes, involved in photosynthesis, hydroxyproline-rich glycoproteins (essential components of plant cell-wall), and meiosis genes. These were unexpected since Mamiellales are generally considered to be asexual, suggesting that these genes were remnants of their common ancestor with land plants, or alternatively that they possessed a kind of sexuality that has not been described yet. This last hypothesis would be compatible with the presence of glycoproteins known to be expressed after sexual fusion in *Chlamydomonas reinhardtii*. In addition to core genes, 14% of proteins (1384) were shared by both *Micromonas* isolates but were not found in Ostreococcus. These include enzymes for plastid peptydoglycan synthesis. These “shared” genes were found to evolve more rapidly than core genes. A large proportion of genes present in only one of the two *Micromonas* isolates exhibited homology to animal or bacterial lineages, supporting their acquisition by horizontal transfer. Altogether, 793 and 826 genes were unique to each of the two *Micromonas* isolates, 689 were specific of *O. tauri* and 249 were unique to *O. lucimarinus*. These variable genes when added to the 7137 core genes and to the 2824 genes shared by at least two of the four genomes, gave a Mamiellales pangenome size of 12,518 genes (Nordberg et al. 2014; Worden et al. 2009).

7 Animal Pangenomes

7.1 Drosophila *Genomes*

Drosophila melanogaster is one of the most intensively studied animal models. The first draft of its genome was published in 2000 (Adams et al. 2000). Its euchromatin part covered ≈120 Mb and contained 13,600 genes, only twice as many as budding yeasts. Following this pioneering work, 11 other fly species originating from Africa, Asia, the Americas, and the Pacific islands were sequenced and compared to *D. melanogaster* reference genome. Gene numbers range from 13,733 for *D. melanogaster*[7] to 17,325 for *Drosophila persimilis*. Sequence comparisons established that 49% of *D. melanogaster* genes were conserved as single-copy orthologues across the whole set of species, defining a set of 6698 core genes. Collectively, the 12 *Drosophila* genomes contain 40,852 variable genes, for a pangenome size of 47,550 genes, but unfortunately it was not possible to determine if this pangenome was closed or open with published data. However, some interesting observations were made. First, effector proteins (like antimicrobial peptides) evolved by rapid duplications and deletions and were significantly underrepresented in the core-genome. Second, gene families forming most of the variable gene content expanded or contracted at a rate of one fixed gene gain or loss every 60,000 years. Common functions among some of the rapidly evolving families include defense response and proteolysis. Third, the vast majority (98%) of *Drosophila* proteins were ancestrally present at the root of the genus. Out of the 296 non-ancestral proteins, 252 were specific of the *Sophophora* subgenus or were complex acquisitions. The remaining 44 genes were lineage-specific (four of them are found only in *D. melanogaster*), were shorter than the average, harbored fewer introns and 40% of them (18/44) were testis-specific, consistent with previous observations about new *Drosophila* genes (Drosophila 12 Genomes Consortium 2007).

In conclusion, *Drosophila* core-genes represent roughly 40–50% of each species gene pool and variable genes arise most of the time by duplication or deletion of an existing gene, with very little de novo gene creation.

7.2 Avian Genomes

Birds encompass the richest variety of species among tetrapod vertebrates, with more than 10,000 different species. In an international effort, 48 avian species, covering most avian clades were sequenced to low or high coverage and compared to the existing three reference genomes (zebra finch, turkey, and chicken), as well as to three crocodilian genomes, the closest bird relatives. After filtering for transposable elements, each genome was predicted to contain ≈14,000–17,000 genes. They

[7] Gene number was refined since publication of the draft genome sequence.

contained a low level of repeated elements (4–10%) as compared to other tetrapods (34–52% in mammals, for example).

Genes responsible for morphological and physiological peculiarities of the clade were analyzed more in depth. Flight capacity was permitted through duplication and positive selection of genes regulating skeleton morphology and bone development. Out of 89 genes involved in ossification half of them showed traces of positive selection, compared to one-third of the 31 orthologous genes in mammals.

Feathers are made of α- and β-keratins, the latter only found in birds and reptiles. Aves genomes contained fewer α-keratin genes as compared to mammals but the repertoire of β-keratins has expanded (up to ≈150 copies in zebra finch). Similarly, most avian genomes contained a higher number of opsin genes than mammalian genomes, partly explaining their more advanced visual system. Genomic elements that were highly conserved among the 48 bird genomes were identified genome wide. Such elements covered 11 Mb (1% of the avian genome) and were significantly underrepresented in coding regions. Actually, the proportion of conserved elements in noncoding regions were 50-fold higher and mostly corresponded to regulatory regions of developmental genes. This result suggested that few avian-specific genes arose in this clade, most of the genomic changes resulting from differences in developmental regulations (Seki et al. 2017).

In conclusion, avian genomes are smaller than mammalian genomes, both in size and in number of genes, due to extensive deletions of chromosomal segments in the ancestral lineage. More precise analyses are now required to sort out core genes from dispensable ones in order to be able to define core- and pangenome sizes and contents.

7.3 Human Genomes

The first human genome drafts were published in 2001 at the same time by the Human Genome Sequencing Consortium and by Celera Genomics (International Human Genome Sequencing Consortium 2001; Venter et al. 2001), and a more complete version of the academic sequence was released in 2004 (International Human Genome Sequencing Consortium 2004). A few years later, James Watson's own genome was deciphered (Wheeler et al. 2008), rapidly followed by the first Asian genome (Wang et al. 2008) and the first African genome (McKernan et al. 2009). The human pangenome was built from comparisons between the NCBI human reference genome and four genomes: Venter's (Celera Genomics), Watson's, YH (Asian genome), and NA18507 (African genome), as well as individual human sequences retrieved from GenBank. Four types of sequence variants were detected: (1) sequences that were frequent in African populations but rapidly declined out of Africa; (2) sequences that were rare in African populations but became more frequent with geographical distance; (3) sequences that were present at a low frequency in European populations; and (4) sequences that were rare in Asian populations. This analysis led to the conclusion that the human pangenome should include 19–40 Mb of additional sequence in addition to the reference genome and that complete coverage of all gene variants should be achieved with the sequencing of 100–150 randomly sampled individuals, worldwide. Analysis of sequences that could not map to the reference genome showed that some of the most abundant genes were those encoding *DUX* homeobox proteins (113 hits in YH and 58 in NA18507), known to be associated with chromatin. Also very frequent were gene families known to be rapidly evolving, such as mucins, zinc-finger proteins, and olfactory receptor proteins (Li et al. 2010).

In conclusion, the present-day human pangenome is still open and will require many more finished sequences in order to be resolved. No doubt that recent efforts to sequence 1070 Japanese genomes (Nagasaki et al. 2015), 2504 individuals from 26 worldwide origins (The 1000 Genomes Project Consortium 2015) or 10,545 human genomes representative of the main human populations (Telenti et al. 2016) should allow to more precisely define human core- and pangenomes and definitely solve this question.

7.4 Reaching for the Metazoan Pangenome

With a wealth of more than 300 metazoan genomes sequenced, defining a core- and a pangenome for multicellular animals could seem a reachable goal. However, with an estimation time for the last common ancestor of all metazoans around 800 million years ago (Erwin et al. 2011), identification of a reliable set of core genes might prove challenging. The sponge *Amphimedon queenslandica* is an early metazoan (Fig. 1) whose genome was sequenced in 2010. It is predicted to contain 18,693 protein-coding genes. Comparison with 4670 metazoan gene families defined a set of 1286 proteins that seem to be metazoan specific, thus defining a draft core-genome for multicellular animals (Srivastava et al. 2010). Many gene expansions observed in the metazoan lineage arose by subsequent tandem or local gene duplications, but extensive work is now needed in order to extract this information from available metazoan genome sequences.

8 The Oceanic Pangenome

The TARA ocean program aims at sampling all planktonic lifeforms of the world's ocean (de Vargas et al. 2015). Metatranscriptomes were established from high-coverage polyA RNA-Seq performed on 441 size-fractionated planktonic communities. Subsequent clustering created a nonredundant set of 116 million transcribed sequences, at least 150 bases long. Despite the sampling effort, it was calculated that 166–190 million sequences would be needed to reach saturation of all oceanic eukaryotic expressed sequences. Half of these sequences had no match in public databases, suggesting that they may correspond to new genes, but most of these (60%) were present as single copies. Transcription of these new genes showed that they were expressed to the same level as known families, suggesting that they were conserved in a smaller number of species or that they were present in less abundant taxonomic groups. Increasing the sampling effort should solve this issue (Carradec et al. 2018). These data, although preliminary and not totally exhaustive, demonstrated that it was possible to extract thousands of new eukaryotic genes belonging to yet uncharacterized species from large oceanic metagenomes. It would be difficult to use the same approach for land eukaryotes for which a comprehensive sampling will be much more tedious and time consuming.

8.1 The Haptophyte Alga Emiliania huxleyi

Marine phytoplankton is responsible for carbon fixation and export to the sea floor as calcite, as well as carbon dioxide release during the calcification process. Their influence on carbon metabolism and export to the deep ocean is complex and crucial for the Earth ecosystem. The haptophyte *E. huxleyi* CCMP1516 reference genome was determined, as well as 13 other isolate genomes, from subarctic to tropical oceanic origins (Read et al. 2013). Repetitive elements were extremely abundant, representing about two-thirds of the sequence and include retrotransposons (1%), DNA transposons (3%), rDNA-related repeats (3%), paralogous genes (10%), tandem repeats and low complexity regions, especially 10–11 bp tandemly repeated minisatellites (34%) and unclassified repeats (16%). These repetitive elements account for a large part of the considerable genome size variability, that ranges from 99 to 133 Mb between isolates (141.7 Mb for the CCMP1516 reference). The reference genome gene content was then compared to three isolates of very distant origins.[8] Out of 30,569 predicted genes in the reference, a total of 5218 (17%) were absent from at least one of three isolates and 364 were missing from all three. Further comparisons with the other isolates strengthened this conclusion: the core-genome contained 20,055 genes, about two-thirds of the reference genes, whereas the remaining genes were variable, making *E. huxleyi* pangenome a complex gene repertoire. Besides repeated elements, the genome encodes many iron-binding proteins, 80 in the core-genome and 30 as variable genes. Iron is essential for calcification and photosynthesis and these differences probably reflect ecological disparities among isolates. In addition, the *E. huxleyi* pangenome encodes 700 proteins whose function relies on metal binding: selenium (49 proteins, 20 gene families), zinc (413 proteins), or copper (65 proteins). Finally, the pangenome contains 26 genes involved in vitamin metabolism, but is unable to synthesize vitamins B_1 and B_{12}, restricting *E. huxleyi* to oceanic regions where these are freely available. In conclusion, the large pangenome of this haptophyte is probably necessary to accommodate its ubiquitous distribution in oceans and illustrates physiological and morphological disparities observed among isolates.

[8] English channel, north-eastern pacific ocean and Great Barrier reef.

9 Where Do Eukaryotic Variable Genes Come From?

At the present time, there are six independent origins for novel eukaryotic genes: interspecific hybridizations, whole-genome duplications, segmental duplications, horizontal gene transfer, single gene duplication, and de novo gene creation (Fig. 5).

9.1 Interspecific Hybridizations

The American botanist Edgar Shannon Anderson published in 1949 a book describing interspecific hybridizations between flowering plants and genotype combinations resulting from these crosses (Anderson 1949). Since then, it became widely accepted among botanists that such events were frequent among plants, resulting in frequent transfers of genes from one species to another. Interspecific hybridizations were very common among yeast species too (Morales and Dujon 2012). Modern brewing yeast, *Saccharomyces pastorianus*, is the offspring of two successive hybridizations, an ancestral one between *Saccharomyces uvarum* and an unknown species and a more recent one between the resulting hybrid and *S. cerevisiae* (Nguyen et al. 2011).

Despite these interesting observations, zoologists were stuck with a very conservative notion of species, based on reproductive isolation, i.e., two species were considered as different if the offspring of their mating was sterile. This remarkably conservative thinking did not take into consideration that many natural fertile interspecific animal hybrids were already described: liger (lion and tiger), pizzlies (polar bear and brown bear), Hawaiian duck (mallard/Laysan duck), *Heliconius* butterflies (*Heliconius cydno* and *H. melpomene*) and Darwin's finches, to name only a few (Pennisi 2016). However, this very conservative way of thinking hit a wall when genome-wide sequencing of ancient human DNA demonstrated that modern *Homo sapiens* were the result of at least two interspecific hybridizations. The first one occurred 50,000–80,000 years ago between *Homo neanderthalensis* and ancestral *Homo sapiens*, after their "out of Africa" journey. This resulted in the retention of 1–4% of Neanderthal genes in all modern *Homo sapiens* genomes, except for those of pure African descent (Green et al. 2010). The second hybridization occurred between offsprings of *Homo sapiens* and *Homo neanderthalensis* and a new species of ancestral human, the Denisovan man (named from the cave in the Siberian Altai mountains in which it was discovered). The hallmark of this hybridization can still be seen in present-day Melanesian populations in which 4–6% of genes come from this ancestral Denisovan man (Prüfer et al. 2014). More recently, the same team discovered the remnants of a 13-year-old girl who was the daughter of a Neanderthal mother and of a Denisovan father, demonstrating that these two ancient human populations also hybridized with each other, around 50,000 years ago (Slon et al. 2018).

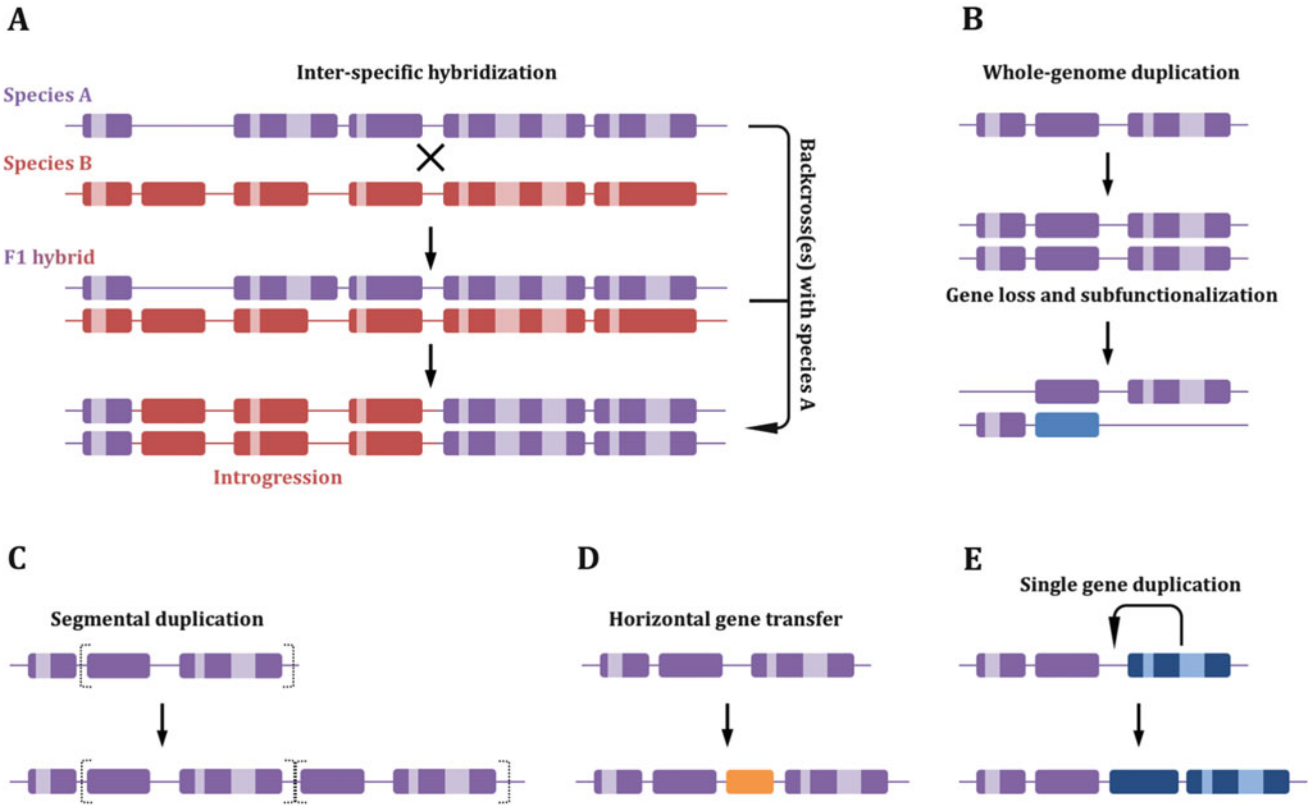


Fig. 5 Gene innovations in eukaryotes. Color boxes represent exons, lighter boxes are introns. (**a**) Interspecific hybridization. Two germ cells with different genomes will merge and produce a fertile hybrid. Several rounds of backcrossing with one of the parents (Species A) will homogenize the genotype but may end

Successive hybridizations can be detected as chromosomal introgressions, large DNA fragments which may be fixed by natural selection following backcrossing between an hybrid and one of its parents (Fig. 5a). One such example in modern humans comes from the Tibetan population. Their genome contains a transcription factor induced under hypoxic conditions, *EPAS1*, whose expression correlates with hemoglobin levels in low atmospheric oxygen pressure. This gene is located in a 120 kb chromosomal region containing a large number of SNPs that were very common in Tibetan and Denisovan DNA, but found at very low frequencies in Han Chinese genomes. This proved that adaptation to high altitude in Tibetan populations was due to a large chromosomal introgression inherited from their Denisovan ancestry (Huerta-Sánchez et al. 2014).

At the present time, it is safe to admit that interspecific hybridizations have been a significant source of gene novelty in eukaryotic genomes, from fungi to animals and plants. However, if living species may mate with other species living in a close ecological niche and produce a fertile offspring, we should now define species independently of the outdated reproductive barrier. Indeed, one may ask what is a species?

9.2 Whole-Genome Duplications

Compared to interspecific hybridizations, bringing together two distinct sets of genes, whole-genome duplications bring together two exact same sets of genes (Fig. 5b). Whole-genome duplications were extremely frequent in every branch of the eukaryote tree, in ascomycetes (Dujon et al. 2004; Kellis et al. 2004; Wolfe and Shields 1997), in paramecium (Aury et al. 2006), in teleostean fish (Jaillon et al. 2004), plants (International Wheat Genome Sequencing Consortium 2014; Jaillon et al. 2007; Vision et al. 2000), rotifers (Flot et al. 2013), and vertebrates (Dehal and Boore 2005), just to cite a few (Fig. 1). These whole-genome duplications were rapidly followed by extensive gene loss, in order to restore gene dosage, but some of the duplicated genes—also called onhologues—may be maintained for a longer time and

Fig. 5 (continued) up in selecting a chromosomal region from the other parent (Species B) that will become a permanent introgression. It is possible that other mechanisms besides backcrossing may generate chromosomal introgressions. (**b**) Whole-genome duplication will be followed by extensive gene loss to counteract gene dosage defects. Sub- or neofunctionalization may occur on one of the two onhologues. Only one chromosome was represented for the sake of clarity, but all chromosomes are duplicated in this process. (**c**) Segmental duplication of a large chromosomal segment (in brackets) may produce several duplicated genes in a single event. (**d**) A gene (in orange) may be transferred from another organism. Horizontal gene transfer may also affect a small number of genes. (**e**) A gene is reversed transcribed and the cDNA integrated in the genome. Former introns are possibly lost in the process if reverse transcription occurs on a spliced transcript. Note that an allelic transposition is represented but ectopic duplications are frequent evolved new functions by neo- or subfunctionalization. *S. cerevisiae* harbors two copies of cytochrome c resulting from an ancestral whole-genome duplication, one encoded by the *CYC1* gene and the other by *CYC7*. The latter is expressed when oxygen levels are so low that cells are in hypoxia, whereas the former is expressed when oxygen levels are normal, a classic case of subfunctionalization (Downie et al. 1977). An interesting example of neofunctionalization was discovered in an Antarctic fish, the eelpout *Lycodichthys dearborni*, whose genome contains two SAS genes, resulting from an ancient duplication. Both SAS-A and SAS-B genes encode an enzyme involved in sialic acid biosynthesis. SAS-B got subsequently partially duplicated and the resulting paralogue was deleted for four out of six exons, making a much shorter gene. The resulting protein happened to bind more efficiently ice crystals than the full-length protein, interfering with crystal growth and behaving as a good antifreeze protein. Subsequent tandem amplifications of this shorter version of SAS-B gave the eelpout the ability to resist extreme cold conditions (Deng et al. 2010).

It might prove technically difficult to discriminate between a recent whole-genome duplication and an interspecific hybridization between two closely related species, without a good reference. It is possible that some chromosomal duplications that were thought to arise from whole-genome duplications were actually acquired by hybridization. In a near future, the achievement of more and more eukaryotic genomes originating from the same clade should eventually dismiss any concern about the origin of close paralogues.

9.3 Segmental Duplications

Another frequent source of novelty comes from local or ectopic duplication of a chromosomal DNA segment, called segmental duplication (Fig. 5c). Their length range from a few to several hundreds of kilobases and they have been found in every eukaryotic species sequenced so far. They are also commonly called copy-number variations (or copy-number variant, or CNV) since their copy number may vary from one genome to another, or structural variant (SV). Spontaneous segmental duplications were found in the yeast *S. cerevisiae*, during experimental evolution of a wild-type strain (Dunham et al. 2002) or using a gene dosage assay for growth recovery (Koszul et al. 2004). These chromosomal duplications could be sometimes quite large, covering 41–655 kb. It was subsequently demonstrated that the mechanism generating segmental duplications was break-induced replication (BIR), a replication-based recombination process that could involve homologous sequences or microhomologies at the junction of duplicated segments (Payen et al. 2008).

Segmental duplications were also described in mouse (Bailey et al. 2004), in primate genomes (Cheng et al. 2005), as well as in man (Bailey et al. 2002). They are known to be associated with several human disorders (Emanuel and Shaikh 2001) and most of them were found to have recently emerged in human history (Jiang et al. 2007). They are undoubtedly a source of gene novelty by successive duplications of large chromosomal segments, although their impact on gene content diversity has not been precisely evaluated yet.

9.4 Horizontal Gene Transfer

Very common between prokaryotes, horizontal gene transfer of a gene (or of a small number of genes) was limited to a few examples in eukaryotes, but may be more widely spread than previously thought. Such events have been identified among *Saccharomycetaceae* yeasts (Fig. 4). Out of 255 species-specific genes, 11 were identified as possible gene transfers from bacterial species, based on sequence similarities and reconstructed phylogenies (Rolland et al. 2009). In *S. pombe*, 34 genes were identified as good candidates for horizontal transfer from bacteria, 16 having occurred before radiation of the clade, 9 being specific to *S. pombe* (Rhind et al. 2011).

Sexuality is a natural obstacle to the propagation of a horizontally acquired gene to metazoan offspring since it must become integrated in the germ line. Nonetheless, some remarkable examples of gene transfer between bacteria or yeast to animal genomes have been described. *Wolbachia pipientis* is a symbiotic bacteria living inside several arthropods and some nematodes. Its genome sequence led to the discovery that 44 out of 45 *Wolbachia* genes were indeed integrated in the genome of the tropical fruit fly *Drosophila ananassae*, one of the natural hosts of this bacteria. Among the other species subsequently screened for the presence of *Wolbachia* genes, one nematode, one mosquito, one tick, three wasps, and five *Drosophila* species contained DNA fragments of various lengths originating from the bacteria (Dunning Hotopp et al. 2007).

Another striking example is the horizontal transfer of yeast genes to pea aphid (*Acyrthosiphon pisum*). This insect displays a red-green color polymorphism that serves to escape its natural predators. The different colors are due to different forms of carotenoid pigments found in individuals. Animals require carotenoids for several essential functions but they are unable to make them. Therefore, they normally find them in their diet. Remarkably, seven carotenoid synthases and carotenoid desaturases, enzymes required for pigment biosynthesis, are encoded by the aphid genome. Comparisons with existing sequences showed that these genes cluster with orthologues from fungi species and subsequent experiments led to the conclusion that these genes were transferred from a fungal pathogen or aphid symbiont, at the root of the aphid clade, followed by subsequent duplications of the transferred gene (s) (Moran and Jarvik 2010).

One last example comes from bdelloid rotifers, near-microscopic animals found in freshwater habitats worldwide. They lost sexual reproduction due to a specific chromosomal organization incompatible with meiotic recombination (Flot et al. 2013). Telomeric regions of *Adineta vaga*, a bdelloid rotifer whose complete genome has been sequenced, revealed dozen of genes of foreign origin. These were found in large telomeric chromosomal segments covering tens of thousands of nucleotides and encoding various proteins playing a role in sugar or amino acid metabolism, in intracellular oxydo-reduction, or in the synthesis of antibiotics and toxins. Most of these genes came from bacteria or fungi species, some of them may have been transferred from plants. Among genes that were identified as of bacterial origin, some harbored introns, whereas their bacterial counterpart did not, suggesting that introns were acquired after transfer from bacteria. Telomeric regions being also enriched in transposable elements, the role of transposons in these massive gene transfers is still an open question (Gladyshev et al. 2008).

9.5 Single-Gene Duplication

Single-gene duplications may occur as allelic or ectopic genome insertions. When occurring in allelic position, they led to tandem repeats of paralogous genes, and were found in variable numbers in eukaryotic genomes. In ascomycetous yeasts, a few dozen tandem gene arrays were detected in each species, mostly composed of two to three copies. However, the *Debaryomyces hansenii* genome contained no less than 247 arrays of tandem paralogues, distributed all over its genome, some of them counting eight or nine tandemly repeated copies (Dujon et al. 2004). Ectopic paralogous gene duplications were also very frequent events in eukaryotes. Most carry the hallmark of retrotransposition: lack of introns, presence of a 3′-end polyA tract and remnants of target site duplications. These retrogenes were also called retroposons (Brosius 1991) and the transposition mechanism was studied in *S. cerevisiae* (Schacherer et al. 2004) as well as in human cells (Esnault et al. 2000). It relies on the reverse transcription of a mature mRNA by a reverse transcriptase (encoded by L1 elements in human cells), followed by integration of the cDNA at an ectopic or allelic locus (Fig. 5e). These duplicated genes lack promoter sequences that were absent from the mature transcript and are therefore pseudogenes, unless they luckily transpose near an active promoter. The human genome contains approximately 10,000 retrogenes, including more than 1700 ribosomal pseudogenes, while the mouse genome contains more than 200 copies of glyceraldehyde-3-phosphate dehydrogenase and *Caenorhabditis elegans* genome harbors more than 2000 pseudogenes (reviewed in Richard et al. 2008).

Extensive retroposition was also frequently detected in plants, the rice genome containing 1235 retrogenes. Interestingly, only 337 (27%) were identified as pseudogenes containing premature stop codons or frameshifts. Subsequent experiments concluded that more than half of the remaining retroposons were probably functional genes. In addition, 380 out of 898 intact retrogenes harbor a chimeric structure containing a flanking exonic sequence (Wang et al. 2006). Therefore, contrarily to the human genome in which most retroposons are pseudogenes, retroposition in the rice genome seems to be an active process rapidly creating new functional genes.

9.6 De Novo Gene Creation

Some remarkable cases of de novo gene invention have been well documented, although the total number of such cases having occurred during evolution of eukaryotes is probably underestimated. *Alu* retrotransposons are very common in primate genomes, being found in more than 1,000,000 copies, covering ≈13% of the genome size and present in almost every protein-coding gene intron (International Human Genome Sequencing Consortium 2001). In dozens of reported cases, an *Alu* sequence was found to be spliced with an upstream exon, resulting in a chimeric peptide (Makałowski et al. 1994). These hybrid proteins are a source of genetic novelty, although their total number in the human genome has not been precisely determined yet.

Before eukaryotes, the living world was asexual, except for bacterial conjugation that may be considered as a very primitive form of mating. Differentiation between two sexes appeared with the first eukaryotic cells and was found almost universally in the eukaryotic world, suggesting that it must be an ancestral acquisition. Sexual reproduction starts with the fusion between two haploid gametes of opposite sex, one male and one female, called syngamy, followed by the merging of both genetical contents. It was recently discovered that the protein responsible for syngamy (called HAP2) was structurally and functionally related to a viral membrane fusion protein. HAP2 was conserved in plants and animals and must have been transferred from a virus to a common ancestor at the root of the eukaryotic lineage (Fédry et al. 2017).

Therian mammals include marsupials and placental (or eutherians), like mouse or man (Fig. 1). In eutherians, egg development takes entire place within the uterus and the placenta is larger and more elaborated than in marsupials. In humans, two genes were responsible for placenta growth, *syncitin-1* and *syncitin-2*. These genes both derived from an envelope protein gene captured from an ancestral virus 25–40 million years ago. Remarkably, the mouse genome harbored two homologues, *syncitin-A* and *syncitin-B*, also deriving from a viral infection in the murine lineage around 20 million years ago, but they are not orthologous to their human counterparts, showing that the placenta was independently invented twice in two mammalian lineages by a similar mechanism of viral gene capture (Dupressoir et al. 2009).

In *D. melanogaster*, the *Sdic* gene coding for a sperm-specific dynein chain was the result of a local duplication and a complex rearrangement between two genes: *Cdic* and *AnnX*. The resulting *Sdic* gene was transcribed from a neo-promoter located in an intronic sequence and the first 21 amino-acids of the resulting protein came from this same intron, now spliced as the first exon of the *Sdic* mRNA (Nurminsky et al. 1998).

One may argue that the above examples are not real de novo gene creations, since they rely on preexisting DNA sequences (*Alu* elements, viral genes, or serendipitous rearrangements of existing exons). It is remarkable that the genome of the excavata *Naegleria gruberi* (Fig. 1) contained 40% of genes without any obvious similarity to any bacterial gene, suggesting that they could be real de novo eukaryotic inventions

(Fritz-Laylin et al. 2010). However, it is possible that many genes that appeared to be novel have indeed diverged so much from their prokaryotic ancestor that they cannot be identified anymore. Hence, the hunt for real de novo gene creation promises to be exciting but seriously challenging!

10 Bioinformatics Tools to Calculate Core- and Pangenomes

Most pangenome analyses were so far performed on prokaryotic genomes. Computing tools rely on the initial determination of genes belonging to the core-genome, followed by addition of all variable genes to build the species pangenome. The initial step is crucial, since one wants to identify the exhaustive list of orthologues belonging to each of the species isolates. Orthologue identification generally uses bidirectional best hits (BDBH), or BLAST followed by a clustering algorithm such as MCL, or comparison of protein domains using Hidden Markov Models (HMM) (reviewed in Guimarães et al. 2015). In a slightly different approach, PanOCT used synteny information in addition to orthology to define the core-genome. The program used a "conserved gene neighborhood" information to discriminate real orthologues from very recently duplicated paralogues whose sequences are indistinguishable (Fouts et al. 2012).

Calculation of eukaryotic core- and pangenomes is significantly more complex for several reasons: (1) the abundance of transposable elements, including novel undescribed transposons absent from dedicated databases; (2) the morcellated nature of genes, particularly in young eukaryotes; (3) the presence of large gene families that make orthologue identification tedious; and (4) the relative incompleteness of genomic sequences, particularly of those containing numerous repeats. In an original approach trying to tackle these problems, genomic and transcriptomic data from 19 *A. thaliana* isolates were analyzed using the GET_HOMOLOGUES-EST software, designed to use tissue-specific expression patterns to build core- and pangenomes. Results support a set of 26,373 core genes and of 11,416 variable genes, for a pangenome containing a total of 37,789 genes. The pangenome is open, each new isolate adding approximately 70 novel variable genes. Core genes exhibit a higher expression level than variable genes and they are under stronger selective pressure ($dN/dS << 1$), confirming what was already observed in other eukaryotes. The same software was used to analyze transcriptomic data from 16 *Hordeum vulgare* isolates (barley), a monocotyledon plant. The barley genome is 34 times larger than *A. thaliana* (4 Gb vs. 119 Mb) and contains 75% of repetitive elements. Its core-genome contains 10,922 genes whereas 28,762 genes were found to be expressed in the leaf transcriptome. Nine isolates were sufficient to sample 99% of the pangenome and its size did not increase with subsequent isolates, proving that it was closed (Table 1). Like *A. thaliana*, core genes were more expressed and more constrained than variable genes (Contreras-Moreira et al. 2017). Merging tissue-specific transcriptomic and whole-genome sequencing data promises to become a powerful approach for future core- and pangenome determinations in metazoans and plants.

11 The Eukaryotic Pangenome

As François Jacob put it more than 40 years ago, gene evolution mainly deals with tinkering, molecular tinkering (Jacob 1977). Young eukaryotes (angiosperms, mammals) reshuffled gene exons and protein domains that already existed in old eukaryotes (fungi, excavata, monocellular animals, and algae), more than one billion years ago. There were very few real inventions after the first eukaryotes, some of them aforementioned here. An *Alu* element or a piece of a virus genome may be captured to make a new protein domain, transposons moved around, sometimes taking along a piece of DNA that would eventually become an exonic sequence, accumulation of mutations in a duplicated gene copy could ultimately create a new function by sub- or neofunctionalization. The redundant nature of eukaryotic genomes, particularly young ones, is only apparent. Eukaryotic core genes are hidden behind legions of transposons, successive whole-genome duplications and interspecific hybridizations, but one may ask how many genes are part of the eukaryotic core-genome. When trying to define it, exons or protein domains, rather than genes, should probably be considered as relevant genetic units, to circumvent issues due to molecular tinkering. Further definition of an eukaryotic pangenome will prove to be a long and complex task, but the accumulation of high-quality genome sequences and the exponential increase of computing power, might prove it to be a reachable goal in the forthcoming years.

In 2016, a German team tried to reconstitute the prokaryotic core-genome, using sequences from 1847 eubacteria and 134 archaebacteria species, covering 6.1 million protein-coding genes belonging to 286,000 families. They identified 355 proteins common to all species, that may be considered as the prokaryotic core-genome (Weiss et al. 2016, 2018). But one may ask whether this minimal set of core genes is sufficient to support life. In an attempt to create a hypothetical minimal genome, the J. Craig Venter institute applied synthetic genomics approaches to *Mycoplasma mycoides*. Using a combination of existing deletion data and literature mining, eight independent segments covering altogether the whole *M. mycoides* genome were synthesized. Each of these eight segments was individually reintroduced into bacteria, but only one of them produced a viable genome. Using high-throughput transposon mutagenesis, the team subsequently identified a set of 229 genes that would cause different levels of growth impairment. The eight DNA segments were rebuilt including these genes. Although each of the individual segment was able to produce a viable genome, addition of the eight segments in the same bacteria was lethal. Once the team eventually solved this synthetic lethality issue and succeeded in synthesizing a fully functional minimal genome, they discovered that the biological function of 146 genes (out of 473 encoded) could not be assigned. These genes of unknown function were all needed to sustain *M. mycoides* life (Hutchison et al. 2016). This interesting work supports the conclusion that designing a minimal genome based on a core set of genes common to several isolates or to several species might not be sufficient to support life. Therefore, defining pangenome contents might prove essential to rewrite the genomes of more complex organisms, like eukaryotes.

As one last word, it must be noted that core-and pangenomes described here took only into consideration protein-coding genes. It is noteworthy that eukaryotes contain many more genes encoding various RNA species: tRNA, rRNA, snoRNA, scRNA, microRNA, and siRNA. Building the whole repertoire of such genes will be challenging but essential to define, at last, a complete eukaryotic pangenome.

Acknowledgments I wish to thank Héloïse Muller for useful suggestions and the Centre National de la Recherche Scientifique (CNRS) as well as the Institut Pasteur for their continuous support.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF et al (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195

Anderson E (1949) Introgressive hybridization. Wiley, New York

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N et al (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444:171–178

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science 297:1003–1007

Bailey JA, Church DM, Ventura M, Rocchi M, Eichler EE (2004) Analysis of segmental duplications and genome assembly in the mouse. Genome Res 14:789–801

Bennett RJ, Johnson AD (2003) Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains. EMBO J 22:2505–2515

Bodey GP, Mardani M, Hanna HA, Boktour M, Abbas J, Girgawy E, Hachem RY, Kontoyiannis DP, Raad I (2002) The epidemiology of *Candida glabrata* and *Candida albicans* fungemia in immunocompromised patients with cancer. Am J Med 112:380–385

Brosius J (1991) Retroposons – seeds of evolution. Science 251:753–753

Brugger J, Feulner G, Petri S (2017) Baby, it's cold outside: climate model simulations of the effects of the asteroid impact at the end of the Cretaceous. Geophys Res Lett 44(1):419–427. https://doi.org/10.1002/2016GL072241

Carradec Q, Pelletier E, Silva CD, Alberti A, Seeleuthner Y, Blanc-Mathieu R, Lima-Mendez G, Rocha F, Tirichine L, Labadie K et al (2018) A global ocean atlas of eukaryotic genes. Nat Commun 9:373

Carreté L, Ksiezopolska E, Pegueroles C, Gómez-Molero E, Saus E, Iraola-Guzmán S, Loska D, Bader O, Fairhead C, Gabaldón T (2018) Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. Curr Biol 28:15–27.e7

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Paabo S et al (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437:88–93

Contreras-Moreira B, Cantalapiedra CP, García-Pereira MJ, Gordon SP, Vogel JP, Igartua E, Casas AM, Vinuesa P (2017) Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. Front Plant Sci 8:184

Cormack BP, Ghori N, Falkow S (1999) An adhesin of the yeast pathogen *Candida glabrata* mediating adherence to human epithelial cells. Science 285:578–582

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Bescot NL, Probert I et al (2015) Eukaryotic plankton diversity in the sunlit ocean. Science 348:1261605

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:1700–1708

Deng C, Cheng C-HC, Ye H, He X, Chen L (2010) Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. PNAS 107:21593–21598

Domergue R, Castaño I, Peñas ADL, Zupancic M, Lockatell V, Hebel JR, Johnson D, Cormack BP (2005) Nicotinic acid limitation regulates silencing of Candida adhesins during UTI. Science 308:866–870

Downie JA, Stewart JW, Brockman N, Schweingruber AM, Sherman F (1977) Structural gene for yeast iso-2-cytochrome c. J Mol Biol 113:369–384

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218

Dujon B (2006) Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. Trends Genet 22:375–387

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E et al (2004) Genome evolution in yeasts. Nature 430:35–44

Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D (2002) Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. PNAS 99:16144–16149

Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G (2012) Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. Genome Res 22:908–924

Dunning Hotopp JC, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Muñoz Torres MC, Giebel JD, Kumar N, Ishmael N, Wang S et al (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317:1753–1756

Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, Heidmann T (2009) Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. PNAS 106:12127–12132

Emanuel BS, Shaikh TH (2001) Segmental duplications: an “expanding” role in genomic instability and disease. Nat Rev Genet 2:791–800

Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. Nature 440:623–630

Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, Peterson KJ (2011) The Cambrian conundrum: early divergence and later ecological success in the early history of animals. Science 334:1091–1097

Esnault C, Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24:363–367

Fabre E, Muller H, Therizols P, Lafontaine I, Dujon B, Fairhead C (2005) Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. Mol Biol Evol 22:856–873

Fédry J, Liu Y, Péhau-Arnaudet G, Pei J, Li W, Tortorici MA, Traincard F, Meola A, Bricogne G, Grishin NV et al (2017) The ancient gamete fusogen HAP2 is a eukaryotic class II fusion protein. Cell 168:904–915.e10

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland, MA

Flot J-F, Hespeels B, Li X, Noel B, Arkhipova I, Danchin EGJ, Hejnol A, Henrissat B, Koszul R, Aury J-M et al (2013) Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. Nature 500:453–457

Fouts DE, Brinkac L, Beck E, Inman J, Sutton G (2012) PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res 40:e172–e172

Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, Kuo A, Paredez A, Chapman J, Pham J et al (2010) The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell 140:631–642

Gabaldón T, Fairhead C (2019) Genomes shed light on the secret life of *Candida glabrata*: not so asexual, not so commensal. Curr Genet 65(1):93–98

Gabaldon T, Martin T, Marcet-Houben M, Durrens P, Bolotin-Fukuhara M, Lespinet O, Arnaise S, Boisnard S, Aguileta G, Atanasova R et al (2013) Comparative genomics of emerging pathogens in the *Candida glabrata* clade. BMC Genomics 14:623

Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered 100:659–674

Gladyshev EA, Meselson M, Arkhipova IR (2008) Massive horizontal gene transfer in bdelloid rotifers. Science 320:1210–1213

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M et al (1996) Life with 6000 genes. Science 274:546–567

Goffeau A, Aert R, Agostini-Carbone ML, Ahmed A, Aigle M, Alberghina L, Albermann K, Albers M, Aldea M, Alexandraki D et al (1997) The yeast genome directory. Nature 387 (suppl):1–105

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK, Severn-Ellis A, McCombie WR, Parkin IAP et al (2016) The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun 7:13390

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y et al (2010) A draft sequence of the neandertal genome. Science 328:710–722

Guimarães LC, Florczak-Wyspianska J, de Jesus LB, Viana MVC, Silva A, Ramos RTJ, Soares S d C, Soares S d C (2015) Inside the pan-genome – methods and software overview. Curr Genomics 16:245–252

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME et al (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716

Hirakawa MP, Martinez DA, Sakthikumar S, Anderson MZ, Berlin A, Gujja S, Zeng Q, Zisson E, Wang JM, Greenberg JM et al (2015) Genetic and phenotypic intra-species variation in *Candida albicans*. Genome Res 25:413–425

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K et al (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26:121–135

Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M et al (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512:194–197

Hutchison CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L et al (2016) Design and synthesis of a minimal bacterial genome. Science 351:aad6253–aad6253

i5K Consortium (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered 104:595–600

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

International Wheat Genome Sequencing Consortium (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345:1251788

Jackson AP, Gamble JA, Yeomans T, Moran GP, Saunders D, Harris D, Aslett M, Barrell JF, Butler G, Citiulo F et al (2009) Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. Genome Res 19:2231–2244

Jacob F (1977) Evolution and tinkering. Science 196:1161–1166

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A et al (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431:946–957

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet 39:1361–1368

Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT et al (2004) The diploid genome sequence of *Candida albicans*. Proc Natl Acad Sci USA 101:7329–7334

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617–624

Koszul R, Caburet S, Dujon B, Fischer G (2004) Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. EMBO J 23:234–243

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J et al (2010) Building the sequence map of the human pan-genome. Nat Biotechnol 28:57–63

Li Y, Zhou G, Ma J, Jiang W, Jin L, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L et al (2014) *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32:1045–1052

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJT, van Oudenaarden A, Barton DBH, Bailes E, Nguyen AN, Jones M, Quail MA, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ (2009) Population genomics of domestic and wild yeasts. Nature 458 (7236):337–341

López-García P, Moreira D (2006) Selective forces for the origin of the eukaryotic nucleus. BioEssays 28:525–533

Makałowski W, Mitchell GA, Labuda D (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet 10:188–193

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC et al (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19:1527–1541

Mereschowsky K (1999) On the nature and origin of chromatophores in the plant kingdom. Eur J Phycol 34:287–295

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on earth and in the ocean? PLoS Biol 9:e1001127

Morales L, Dujon B (2012) Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. Microbiol Mol Biol Rev 76:721–739

Moran NA, Jarvik T (2010) Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science 328:624–627

Muller H, Thierry A, Coppée J-Y, Gouyette C, Hennequin C, Sismeiro O, Talla E, Dujon B, Fairhead C (2009) Genomic polymorphism in the population of *Candida glabrata*: gene copy-number variation and chromosomal translocations. Fungal Genet Biol 46(3):264–267. https://doi.org/10.1016/j.fgb.2008.11.006

Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S et al (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun 6:8018

Nguyen H-V, Legras J-L, Neuvéglise C, Gaillardin C (2011) Deciphering the hybridisation history leading to the lager lineage based on the mosaic genomes of *Saccharomyces bayanus* strains NBRC1948 and CBS380T. PLoS One 6:e25821

Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I (2014) The genome portal of the department of energy joint genome institute: 2014 updates. Nucleic Acids Res 42:D26–D31

Normile D, 2017, and Am, 8:00 (2017) Plant scientists plan massive effort to sequence 10,000 genomes

Nurminsky DI, Nurminskaya MV, Aguiar DD, Hartl DL (1998) Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 396:572–575

Parkin IA, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL et al (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. Genome Biol 15:R77

Payen C, Koszul R, Dujon B, Fischer G (2008) Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms. PLoS Genet 5: e1000175

Pennisi E (2016) Shaking up the tree of life. Science 354:817–821

Pennisi E (2017) Biologists propose to sequence the DNA of all life on Earth. Science Magazine, Feb 27, 2017

Peter J, Chiara MD, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A et al (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature 556:339–344

Pfaller MA, Diekema DJ (2004) Twelve years of fluconazole in clinical practice: global trends in species distribution and fluconazole susceptibility of bloodstream isolates. Clin Microbiol Infect 10:11–23

Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC, Zaina G, Bastien C, Cattonaro F, Marroni F et al (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. Mol Biol Evol 33:2706–2719

Polakova S, Blume C, Zarate JA, Mentel M, Jorck-Ramberg D, Stenderup J, Piskur J (2009) Formation of new chromosomes as a virulence mechanism in yeast *Candida glabrata*. Proc Natl Acad Sci USA 106:2688–2693

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C et al (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49

Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier A, Salamov A et al (2013) Pan genome of the phytoplankton *Emiliania* underpins its global distribution. Nature 499:209–213

Renne PR, Sprain CJ, Richards MA, Self S, Vanderkluysen L, Pande K (2015) State shift in Deccan volcanism at the Cretaceous-Paleogene boundary, possibly induced by impact. Science 350:76–78

Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, Wapinski I, Roy S, Lin MF, Heiman DI et al (2011) Comparative functional genomics of the fission yeasts. Science 332:930–936

Richard G-F, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72:686–727

Rolland T, Neuvéglise C, Sacerdot C, Dujon B (2009) Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. PLoS One 4:e6515

Rolland T, Dujon B, Richard GF (2010) Dynamic evolution of megasatellites in yeasts. Nucleic Acids Res 38:4731–4739

Sagan L (1967) On the origin of mitosing cells. J Theor Biol 14:225–IN6

Schacherer J, Tourrette Y, Souciet J-L, Potier S, de Montigny J (2004) Recovery of a function involving gene duplication by retroposition in *Saccharomyces cerevisiae*. Genome Res 14:1291–1297

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Seki R, Li C, Fang Q, Hayashi S, Egawa S, Hu J, Xu L, Pan H, Kondo M, Sato T et al (2017) Functional roles of Aves class-specific *cis*-regulatory elements on macroevolution of bird-specific features. Nat Commun 8:14229

Slon V, Mafessoni F, Vernot B, de Filippo C, Grote S, Viola B, Hajdinjak M, Peyrégne S, Nagel S, Brown S et al (2018) The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature 561:113–116

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U et al (2010) The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466:720–726

Tekaia F, Dujon B, Richard G-F (2013) Detection and characterization of megasatellites in orthologous and nonorthologous genes of 21 fungal genomes. Eukaryot Cell 12:794–803

Telenti A, Pierce LCT, Biggs WH, di Iulio J, Wong EHM, Fabani MM, Kirkness EF, Moustafa A, Shah N, Xie C et al (2016) Deep sequencing of 10,000 human genomes. PNAS 113:11901–11906

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". PNAS 102:13950–13955

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11:472–477

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74

The 1001 Genomes Consortium (2016) 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell 166:481–491

Thierry A, Bouchier C, Dujon B, Richard G-F (2008) Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in *Candida glabrata*. Nuclic Acids Res 36:5970–5982

Thierry A, Dujon B, Richard G-F (2009) Megasatellites: a new class of large tandem repeats discovered in the pathogenic yeast *Candida glabrata*. Cell Mol Life Sci 67:671–676

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al (2001) The sequence of the human genome. Science 291:1304–1351

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. Science 290:2114–2117

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S et al (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18:1791–1802

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J et al (2008) The diploid genome sequence of an Asian individual. Nature 456:60

Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF (2016) The physiology and habitat of the last universal common ancestor. Nat Microbiol 1:16116

Weiss MC, Preiner M, Xavier JC, Zimorski V, Martin WF (2018) The last universal common ancestor between ancient Earth chemistry and the onset of genetics. PLoS Genet 14:e1007518

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713

Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV et al (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science 324:268–272

Yao W, Li G, Zhao H, Wang G, Lian X, Xie W (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol 16:187

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU et al (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358

Zhang G (2015) Genomics: bird sequencing project takes off. Nature 522:34

Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T et al (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50:278–284

Computational Strategies for Eukaryotic Pangenome Analyses

Zhiqiang Hu, Chaochun Wei, and Zhikang Li

Abstract Over the last few years, pangenome analyses have been applied to eukaryotes, especially to important crops. A handful of eukaryotic pangenome studies have demonstrated widespread variation in gene presence/absence among plant species and its implications on agronomically important traits. In this chapter, we focus on the methodology of pangenome analysis, which can generally be classified into two different types of approaches, a homolog-based strategy and a "map-to-pan" strategy. In a homolog-based strategy, the genomes of individuals are independently assembled, and the presence/absence of a gene family is determined by clustering protein sequences into homologs. Alternatively, in a "map-to-pan" strategy, pangenome sequences are constructed by combining a well-annotated reference genome with newly identified non-reference representative sequences, from which the presence/absence of a gene is then determined based on read coverage after individual reads are mapped to the pangenome. We highlight the advantages and limitations of the homolog-based strategy and several variant approaches to the "map-to-pan" strategy. We conclude that the "map-to-pan" strategy is highly recommended for eukaryotic pangenome analysis. However, programs and parameters for pangenome analysis need to be carefully selected for eukaryotes with different genome sizes.

Z. Hu (✉)
Department of Plant & Microbial Biology, University of California, Berkeley, CA, USA

Institute of Crop Sciences/National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Haidian District, Beijing, China
e-mail: hu.zhiqiang@berkeley.edu

C. Wei
School of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

Z. Li
Institute of Crop Sciences/National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Haidian District, Beijing, China

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_13

Keywords Pangenome · Plant pangenome · Gene presence–absence variation · Gene PAV · Map-to-pan

In 2005, Tettelin et al. introduced the concept of a pangenome, namely the entire gene set of a species, in their study of eight strains of *Streptococcus agalactiae*, that causes neonatal infection in humans (Tettelin et al. 2005). The pangenome is comprised of a "core-genome" that contains genes shared by all individuals of the species, and a "dispensable genome" containing genes present in some but not all individuals of the species. The core-genome is generally believed to be responsible for functions essential to the species, such as growth and development, whereas the dispensable genome confers functions related to environmental adaptations (Vernikos et al. 2015). During the past 10 years, pangenome studies have been widely applied to bacteria and other microorganisms. However, only a handful of pangenome analyses of higher eukaryotes have been reported (Wang et al. 2018; Hu et al. 2018; Sun et al. 2017; Zhao et al. 2018; Ou et al. 2018; Darracq et al. 2018; Montenegro et al. 2017; Pinosio et al. 2016; Golicz et al. 2016; Nguyen et al. 2015; Lu et al. 2015; Yao et al. 2015; Hirsch et al. 2014; Read et al. 2013; Li et al. 2010, 2014). In this chapter, we will first review the biological insights highlighted from these studies. Then, we will introduce current challenges and strategies for performing eukaryotic pangenome analysis, and finally, we will discuss future directions in this field.

Next-generation sequencing (NGS) technologies have enabled whole-genome sequencing and comparisons of multiple individual genomes within a species. Single nucleotide variations (SNPs), small insertions and deletions (InDels), and structural variations (SVs), including copy number variations (CNVs) and presence/absence variations (PAVs), can be identified when comparing against a reference genome. A considerable number of SVs have been observed among human (Sudmant et al. 2015; Genomes Project et al. 2015; Feuk et al. 2006) and animal genomes (Bickhart and Liu 2014). For example, a typical human genome contains 2100–2500 structural variants (including ~1000 large deletions), affecting ~20 Mb sequences when comparing with a reference genome (~3 Gb) (Genomes Project et al. 2015). In contrast, SVs have been reported to be more pervasive within plant genomes (Saxena et al. 2014), such as rice (Wang et al. 2018; Hu et al. 2018), arabidopsis (Cao et al. 2011), maize (Swanson-Wagner et al. 2010), sorghum (Zheng et al. 2011), and potato (Potato Genome Sequencing C et al. 2011). For example, the total sequences affected by SV that differentiate two typical rice accessions, on average, are about 22–70 M (out of ~380 M) (Wang et al. 2018). These results imply that there might be widespread presence of gene PAVs associated with SV sequences.

Pangenome analyses aim to study gene PAVs, providing a new functional interpretation of within-species variations. Compared to SV studies, pangenome analyses identify undiscovered genomic sequences and their associated genes and reveal the species core and dispensable genome. Early pangenome studies focused on comparisons among a small number (2–3) of well-assembled individual genomes

(Liu et al. 2007; Ma and Bennetzen 2004). These studies revealed the space of undiscovered genes and demonstrated widespread gene PAVs within a species. For instance, Li et al. assembled an Asian and an African genome, leading to the detection of 5 Mb sequences and hundreds of undiscovered genes that are absent in the human reference genome. Liu et al. sequenced ten thousand cDNAs of 93–11, a *Xian*(*indica)* rice accession, and found that >1000 genes were absent in the *Geng (japonica)* reference genome (Liu et al. 2007), which was believed to have diverged from *Xian* ~0.44 million years ago (Ma and Bennetzen 2004); later, Schatz et al. compared three assembled genomes of a *Xian* (IR64), a *Geng,* and an *aus* (DJ123) accession, and found that ~3000 genes were absent in at least one accession.

However, studying a small number of individuals cannot reveal the global landscape of gene PAVs of a species and cannot confidently identify the species core and dispensable genomes. Thus, systematic studies involving a large number of representative individuals within a species is highly desired. Large-scale plant pangenome studies involving tens to hundreds of individuals have emerged over recent years (Table 1). Many of these studies revealed that gene PAVs are a very important aspect of the genomic diversity within eukaryotic species/populations that can provide significant insights into evolutionary history of the species/populations with significant implications on the functional genomic research of important traits.

In *Emiliania huxleyi*, a marine phytoplankton important for carbon fixation in ecosystems, one-third of the genes in the reference genome are absent in the 13 sequenced individuals (Read et al. 2013). The core-genome controls inorganic nitrogen uptake/assimilation and nitrogen-rich compound acquisition/degradation, while the dispensable genome is in charge of metabolic repertoires, of which over one-fourth involve iron-binding activities and vitamin B1 and B12 synthesis (Read et al. 2013).

In rice, several studies consistently report that about ten thousand genes are missing in the widely used Nipponbare reference genome (Wang et al. 2018; Zhao et al. 2018; Yao et al. 2015), and almost all of them can be detected in wild rice (Wang et al. 2018). The dispensable genome accounts for >38% of the species pangenome and over one-fourth of a typical individual genome (Wang et al. 2018). On average, two *Xian* or *Geng* genomes differ by about 4000 (~10%) genes, respectively, whereas a *Xian* genome and a *Geng* genome differ by more than 6000 (~15%) genes (Wang et al. 2018). Although the dispensable genome is less studied, it appeared to harbor functions related to environmental adaptations, such as regulation of immune/defense responses and ethylene metabolism (Wang et al. 2018). Interestingly, the well-known Green Revolution gene, *sd-1*, coding a key enzyme, GA-oxidase20, in the biosynthesis of the important plant hormone, GA_1/GA_4, is a dispensable gene that associates with many important processes in plant growth, development, and responses to abiotic stresses (Wang et al. 2018; Zhao et al. 2018).

In *Brassica oleracea* (Golicz et al. 2016), bread wheat (Montenegro et al. 2017) and wild soybean (Li et al. 2014), it was reported that the dispensable genomes take up 18.7%, 20%, and 35.7% of the pangenomes, respectively. Although the pipelines and parameters/thresholds used to determine gene presence differed a lot in the above studies, it is well demonstrated that plants exhibit considerably large

Table 1 Representative pangenome studies

Species	Haploid genome size (bps)[a]	N	References	Strategy	Comment
Homo sapiens (human)	2991 M	3	Li et al. (2010)	Directly comparing two de novo assembled individual human genomes (an Asian and an African) with the human reference genome.	19~40 Mb sequences containing >150 genes cannot be found in the reference.
Emiliania huxleyi (coccolithophore)	168 M	14	Read et al. (2013)	Building a reference genome from an individual genome; assembling 3 additional individual genomes and comparing them with the reference genome; determining presence/absence of reference genes by mapping short reads of additional 10 individuals to the reference.	>1300 reference genes are not present in the 3 individual genomes; the core-genome accounts for 2/3 of the reference genes.
Zea mays (maize)	2135 M	503	Hirsch et al. (2014)	Sequencing the transcriptome of 503 accessions. Assembling genes from transcriptome sequencing. A gene with FPKM > 0 is considered as present.	Identifying ~8600 representative transcript assemblies (RTAs) absent in the B73 reference; 16.4% RTAs express in all lines and 82.7% express in subsets of the lines.
Glycine soja (wild soybean)	924 M	7	Li et al. (2014)	Sequencing and de novo assembling 7 individuals' genomes. Clustering annotated genes to gene families.	Dispensable genome accounts for 20% of the pangenome, and displays greater sequence variation than the core-genome.
Oryza sativa (rice)	374 M	1483	Yao et al. (2015)	Aligning low-depth (1~3x)	Detecting ~9000 genes for the

(continued)

Table 1 (continued)

Species	Haploid genome size (bps)[a]	N	References	Strategy	Comment
				reads to a pangenome; building the dispensable genome by assembling the pool of unaligned reads from each individual. *Indica* and *japonica* accessions are separately studied.	*indica* dispensable genome and >6000 genes for *japonica* dispensable genome.
Brassica oleracea	514 M	9	Golicz et al. (2016)	Using a reference-based iterative strategy to assemble the pangenome: (1) mapping reads to the reference sequence; (2) assembling unmapped reads; (3) and updating the reference. Determine gene PAV by mapping short reads to the pangenome.	Dispensable genome accounts for 18.7% of the pangenome.
Triticum aestivum (bread wheat)	13,672 M	18	Montenegro et al. (2017)	Building a reference genome; Constructing the pangenome sequences by combining the reference genome and non-reference sequences, which are assembled from the pool of unaligned reads from each individual. Determining gene presence/absence by mapping short reads to the pangenome.	Dispensable genome accounts for 35.7% of the pangenome.

(continued)

Table 1 (continued)

Species	Haploid genome size (bps)[a]	N	References	Strategy	Comment
Oryza sativa (rice)	374 M	453	Wang et al. (2018), Hu et al. (2018), Sun et al. (2017)	Assembling 3010 individual genomes independently; building representative non-reference sequences by removing the redundant sequences from the pool of contigs that are unaligned to the reference. Constructing a pangenome by combining the reference genome and representative non-reference sequences. Determining gene presence/absence for 453 individuals with sequencing depth >20 by mapping short reads to the pangenome.	Discover 283 M non-reference sequences with >10,000 genes; Dispensable genome accounts for 35.7% of the pangenome. Dispensable genes tend to be younger, shorter, exhibiting higher level of SNPs.
Capsicum (including 4 species) (pepper)	3095 M	383	Ou et al. (2018)	Using the same strategy as the above rice study.	Discover 956 M non-reference sequences with >50,000 genes; 55.7% of the pangenome show >50% presence frequencies in all the 4 species.
Oryza sativa and *Oryza rufipogon* (rice and wild rice)	374 M	66	Zhao et al. (2018)	Sequencing and de novo assembling 66 individual genomes. Clustering annotated genes to gene families.	Discover >10,000 non-reference genes; 62% of the pangenome can be found in ≥60 individuals.

[a] The genome size was obtained from NCBI genome database. It can be the size of a reference genome or the average size of several independent assemblies

dispensable genomes, harboring functions related to many agronomically important traits. Moreover, several studies consistently demonstrate that dispensable genes tend to be younger (Wang et al. 2018; Chen et al. 2012; Bush et al. 2013), shorter (Wang et al. 2018; Bush et al. 2013; Schatz et al. 2014), have less exons (Wang et al. 2018; Bush et al. 2013; Schatz et al. 2014), harbor a much higher level of sequence variations (Wang et al. 2018; Li et al. 2014), and have fewer paralogs (Wang et al. 2018; Bush et al. 2013).

1 Eukaryotic Pangenome Analysis Strategy

Because pangenome is a property of a species/population, any desirable pangenome study should seriously consider its sampling strategy such that the maximum gene PAVs can be detected with a minimum number of samples. According to the core collection concept in plant genetic resources (Frankel and Brown 1984), a core collection of a plant species germplasm consisting of limited but well-sampled accessions of a plant species would represent the whole spectrum of its total within-species diversity. In practice, a well-established semi-stratified sampling strategy considering both the center(s) of diversity/origin and geographic distribution of a plant species has demonstrated that the core collection containing only 5% of the total collected accessions of a species would cover ~95% of the total species diversity (Jia et al. 2017). Obviously, this concept should equally be applicable to pangenome research of animal species.

For the analytic methodology, almost all bacterial pangenome analyses follow a homolog-based strategy (Fig. 1) involving (1) de novo assembly of individual genomes; (2) independent annotation of protein-coding genes in each assembled genome; and (3) pooling all protein sequences together and clustering them into homologs (gene families) or orthologs using protein clustering tools (Steinegger and Söding 2018; Fu et al. 2012) or ortholog grouping tools (Emms and Kelly 2015; Li et al. 2003). Gene family presence/absence in each individual can be retrieved from

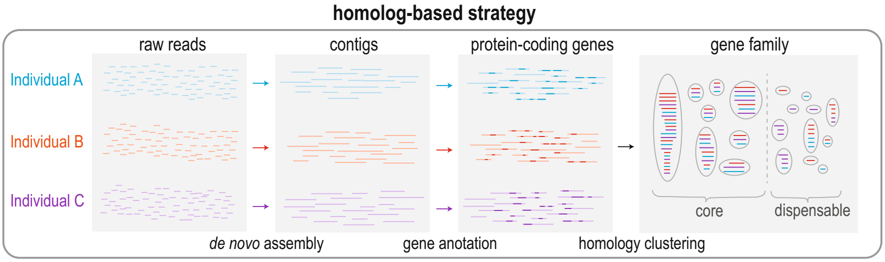


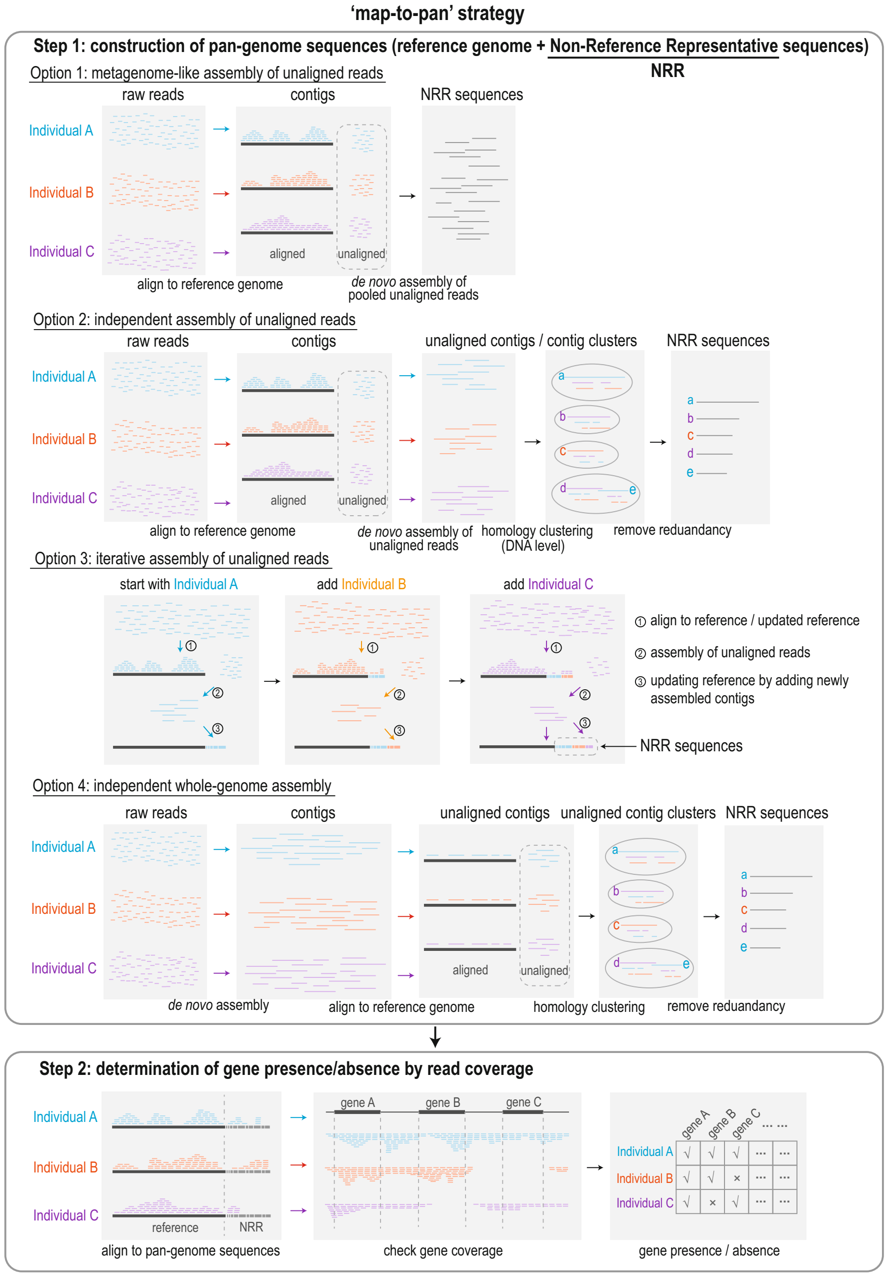
Fig. 1 Homolog-based strategy for pangenome analyses. This strategy is widely used for bacterial pangenome analyses. It includes the following steps: (1) independent assemblies of individual whole genomes; (2) annotation of protein-coding genes for each genome; and (3) clustering genes to homologs (gene families) to determine the presence/absence of each gene family

the clustering results. This strategy is highly dependent on the completeness of the whole-genome assembly. Failure in assembling a sequence segment will lead to calling the absence of all genes located on this sequence segment. Moreover, the protein similarity threshold for gene family determination may impact the size and even the relative size of the core-genome and pangenome.

Several challenges hinder applying a homolog-based strategy to eukaryotic genomes. First, eukaryotic genomes are usually large, ranging from hundred millions of bases to billions of bases, and possess a high level of repetitive sequences, making whole-genome assembly challenging. Several approaches can help improve the assembly, including increasing the sequencing depth, sequencing multiple libraries with diverse insertion sizes, and integrating long-read sequencing technologies (Rhoads and Au 2015; Schneider and Dekker 2012). However, all of these approaches significantly increase the cost of whole-genome assembly, thus limiting the number of individuals involved in a study. Second, eukaryotes have split gene structures. Automatic gene annotation may be inaccurate and lead to biased results. Even with these challenges, there are several studies following the homolog-based strategy. Li et al. sequenced seven wild soybean genomes using Illumina technology, each with three libraries (insertion sizes of 180 bp, 500 bp, and 2000 bp) (Li et al. 2014). The average overall sequencing depth was 112x. Based on this data, they were able to assemble ~89% of the genome. Recently, Zhao et al. sequenced 66 rice and wild rice accessions, each with two libraries (insertion sizes of 400 bp and 700 bp) (Zhao et al. 2018). The average sequencing depth reached 115x, and they were able to assemble ~85% of the genomes. Notably, a significant portion of individual genomes were not assembled in both studies. The associated genes were labeled as "absent" in the corresponding individuals. However, given that these false negatives repeatedly happen for certain genes within repeat-rich regions, they can be treated as systematic errors. The overall results may be still meaningful.

Reference-based genomic studies are prevalent in eukaryotes. Researchers have been taking tremendous efforts to build more complete reference genomes and providing confident gene annotations for important species. These reference genomes and their annotated genes are the basis for modern genomics studies. Moreover, reference-based genomic variants show a great power in explaining phenotypic variations when used as markers for genome-wide association analyses. Therefore, when introducing the pangenome concept to eukaryotic genomic analyses, taking advantage of a pre-existing well-annotated reference genome is a straightforward choice. Following this idea, the "map-to-pan" strategy became prevalent for eukaryotic pangenome studies, especially when the target genome is extremely large or the study involves a large number of individuals (Fig. 2). The "map-to-pan" strategy includes two main steps: construction of pangenome sequences by combining the reference genome and non-reference representative (NRR) sequences (upper panel of Fig. 2) and determination of the presence/absence of each gene in each individual by mapping reads to the pangenome and examining the gene coverage (lower pane of Fig. 2).

Several approaches for detecting NRR sequences have been reported (Wang et al. 2018; Ou et al. 2018; Montenegro et al. 2017; Yao et al. 2015; Read et al. 2013; Li


'map-to-pan' strategy
Step 1: construction of pan-genome sequences (reference genome + Non-Reference Representative sequences)
NRR
Option 1: metagenome-like assembly of unaligned reads
raw reads
contigs
NRR sequences
Individual A
Individual B
Individual C
aligned
unaligned
align to reference genome
de novo assembly of pooled unaligned reads
Option 2: independent assembly of unaligned reads
raw reads
contigs
unaligned contigs / contig clusters
NRR sequences
Individual A
Individual B
Individual C
aligned
unaligned
a
b
c
d
e
align to reference genome
de novo assembly of unaligned reads
homology clustering (DNA level)
remove reduandancy
Option 3: iterative assembly of unaligned reads
start with Individual A
add Individual B
add Individual C
① align to reference / updated reference
② assembly of unaligned reads
③ updating reference by adding newly assembled contigs
NRR sequences
Option 4: independent whole-genome assembly
raw reads
contigs
unaligned contigs
unaligned contig clusters
NRR sequences
Individual A
Individual B
Individual C
aligned
unaligned
a
b
c
d
e
de novo assembly
align to reference genome
homology clustering
remove reduandancy
Step 2: determination of gene presence/absence by read coverage
Individual A
Individual B
Individual C
reference
NRR
gene A
gene B
gene C
align to pan-genome sequences
check gene coverage
gene presence / absence

	gene A	gene B	gene C	...	...
Individual A	√	√	√	...	...
Individual B	√	√	×	...	...
Individual C	√	×	√	...	...

Fig. 2 "Map-to-pan" strategy for pangenome analyses. This strategy is mostly used for eukaryotic pangenome analyses. It includes two main steps: (1) construction of pangenome sequences by integrating a reference genome and assembled non-reference sequences; (2) determination of presence/absence of each gene (both reference genes and non-reference predicted genes) based on read coverage. Four strategies for obtaining non-reference representative sequences are introduced

et al. 2014). Yao et al. utilized a metagenome-like assembly of mixed unaligned reads from 1483 rice accessions with extremely low sequencing depth (1~3x) (Yao et al. 2015) (Option 1 in Fig. 2), enabling the detection of ~9000 non-reference genes. This approach assembled NRR sequences using heterozygous reads and may generate chimeric contigs, especially when considering that non-reference sequences may exhibit higher levels of repetitive sequences. A variant of this option (Option 2 in Fig. 2) is to assemble the unaligned reads from each individual separately and retrieve NRR sequences using DNA homology clustering strategies, such as CD-HIT-EST (Fu et al. 2012), UCLUST (Edgar 2010), MeShClust (James et al. 2018), etc. Golicz et al. utilized an iterative assembly approach (Option 3 in Fig. 2), iteratively conducting the following three steps: mapping of the reads to a pseudo pangenome (starting with the reference genome); assembling the unmapped reads; and updating a new pseudo pangenome with new sequences added (Golicz et al. 2016). They demonstrated that the sizes of final assemblies were similar regardless of the order of individuals added into the iterative process. However, an improper ordering may lead to fragmented assemblies. Alternatively, Hu et al. proposed an integrated approach (implemented in EUPAN toolkit (Hu et al. 2017)) (Option 4 in Fig. 2): (1) independent assembly of individual genomes; (2) generation of NRR sequences from homology clustering of all unaligned contigs. This approach has the benefit of not involving chimeric sequences as well as keeping better sequence completeness. This approach has also been recently applied to hundreds of rice genomes (Wang et al. 2018; Hu et al. 2018; Sun et al. 2017) and the 383 Capsicum genomes (Ou et al. 2018). This strategy will perform better than Option 2 in the scenario where a novel sequence contains a short reference segment (likely to be repetitive sequences) in the middle; option 2 will assemble two segmented segments instead. However, the process of whole-genome assembly is computationally intensive, hindering its application to extremely large genomes. In summary, pooling of low-depth sequenced genomes may also contribute to pangenome construction (Option 1). Options 2–4 are preferable if sequencing depth is high enough for independent assemblies. Options 2–3 are extremely useful for eukaryotes with very large genomes (e.g., the bread wheat with a haploid genome of >13Gb).

After the construction of pangenome sequences, gene presence/absence can be determined by examining gene coverage when raw reads are mapped to the pangenome (lower panel of Fig. 2). Remarkably, very different thresholds have been applied to determine a gene's presence. For example, Wang et al. considered a gene's presence as CDS coverage ($\geq$1 read) over 0.95 and gene body coverage over 0.85 (Wang et al. 2018); Ou et al. treated a gene's presence as CDS coverage ($\geq$1 read) over 0.6 and gene body coverage over 0.5 (Ou et al. 2018); Read et al. considered a gene's presence as gene body coverage ($\geq$1 read) over 0.5 (Read et al. 2013; Montenegro et al. 2017; Golicz et al. 2016) used a threshold of exon coverage over 0.05. Unfortunately, such divergent thresholds make the quantitative cross-species comparisons of gene PAV-related features meaningless. Theoretically, with a high-enough sequencing depth, a gene's presence is equal to that the gene, at least the CDS, should be fully covered. Loss of partial sequences of a gene, defined as a "functional unit," may cause a loss of gene function. Setting up gene body coverage cutoffs will help distinguish retro-transcribed pseudo-genes from their original ancestries. In reality, certain genomic regions may be not covered due to both insufficient sequencing depth and unevenness of the sequencing. One plausible solution is to lower the thresholds. However, the sequencing depth difference may further lead to inconsistencies in sensitivities of gene presence determination among individuals; individuals with higher sequencing depth would contain more genes. Another possible solution is to study the presence/absence of gene families instead of genes by calculating "gene presence" using a low threshold and determining gene family presence based on "gene presence." In this scenario, the unbalanced sequencing depths also need to be fixed either by sampling to equal depths or setting up dynamic thresholds based on the sequencing depth. Nevertheless, it is not recommended to determine gene presence/absence from low-depth sequencing data. Gene presence/absence should only be studied and compared for individuals with sufficient sequencing data, that is, when mapping to the pangenome, the coverage of the genome should be saturated. For example, Wang et al. mapped raw reads of ~3000 rice accessions to the reference genome and found that genome coverage is stable when sequencing depth exceeds 20x; therefore, gene presence/absence was only studied for a selected set of 453 accessions with sequencing depth >20 (Wang et al. 2018).

The "map-to-pan" strategy also exhibits better accuracy. A pangenome study can be technically evaluated at two levels: (1) the accuracy of pangenome (gene annotation and gene completeness) and (2) the accuracy of gene presence/absence calling. The "map-to-pan" strategy utilizes reference sequences and their annotations directly. Strategies using a whole-genome assembly (homolog-based, and option 4 of the "map-to-pan" strategy) will have a higher possibility of detecting complete gene sequences. At the gene presence/absence level, the homolog-based strategy has a bottleneck in assembling a complete genome, and "map-to-pan" strategies definitely show better accuracy when sequencing depth is high enough (Hu et al. 2017).

After determination of gene presence/absence, similar analyses as seen in bacterial pangenome studies can be performed for eukaryotes, including but not limited to (1) simulating the pangenome and core-genome sizes; (2) constructing phylogenic relationships based on gene presence/absence; and (3) exploring functions related to the dispensable genome or to a specific dispensable gene.

2 Future Directions

In summary, the pangenome is an important property of any eukaryotic species/populations and gene PAVs represent a very important dimension of within-species/population diversity that remains uncharacterized in most eukaryotic species. As the costs in genome sequencing decrease, one would expect the pangenome analyses to be carried out in more and more species, firstly in most important and/or model plant and animal species, and then to natural populations of wild species. Thus, eukaryotic pangenome research in the next several years should focus on revealing within-species/population gene PAVs and building the pan-references for species of interest. The pan-reference of a species should include the reference illustrating (1) all the sequences within the species, (2) the connections of alternative sequence segments and (3) the genotype likelihoods (allele frequencies) such that all possible mechanisms (SVs and distribution/activities of transposable elements) potentially responsible for pangenome expansion and generation of gene PAVs can be clearly represented and understood. As pangenomes and gene PAVs are revealed in more and more plant and animal species, the eukaryotic pangenome research will be naturally extended to the comparative pangenome analyses, focusing on comparisons of the pangenome constitution between or among related species. Results from this kind of research are expected to provide new insights into the evolutionary history of eukaryotic species. For example, comparisons between related species or between different populations of the same species in portions of the core and dispensable genes/gene families in their pangenomes and their patterns how new gene emerged will provide important information on their evolutionary history. Expectedly, emergences of new species would be accompanied with bursts of new gene emergences, while major distinctions with massive gene losses in evolution. Also, it would be of great interest to compare the core-genome constitution between related species and to compare the dispensable genome constitution between different populations of the same species. In the former cases, one may see the differences in key genes and their functionalities between related species. In the latter cases, one may discover important sets of genes contributing to adaptations to specific environments important for future plant and animal improvements. In this respect, genome-wide association analyses of important traits based on pangenome SNPs or based on gene PAVs should be widely adopted (Hu et al. 2018).

As more eukaryotic pangenome analyses are expected to emerge, the technical strategy and methodology in analyses of eukaryotic pangenomes need to be improved. Because of the relatively high genome sequencing and analytic costs in eukaryotic pangenomes, the NGS technology will remain the primary technology for the pangenome studies of most eukaryotes in the short term, particularly for those species of very large genomes, and so for the "map-to-pan" strategy elaborated in detail here. However, before applying this strategy, specific attentions should be paid to the sampling strategy to make sure representative individuals of minimum sample size of the target species or population to be used, and to the selection and evaluation of parameters of the map-to-pan methodology. In the presentation and storage of results from the eukaryotic pangenome analyses, graph-based data structures are highly desirable and should be widely used in pan-reference storage and visualization (Zekic et al. 2018; Marschall et al. 2018; Baier et al. 2016). Pioneer work has been done in the human genome research, where the NRR sequences might be of a small size. Alternative sequences of highly variable regions were added to human reference genome, starting with GRCh37 (Church et al. 2011). Alternative sequences were anchored to locations along the primary assembly. Besides the limited NRR sequences, a large number of SNPs, InDels, and SVs (deletions, duplications, and translocations) can also be integrated into the pan-reference (Zekic et al. 2018; Marschall et al. 2018; Baier et al. 2016). What is more, read alignment tools and variant-calling tools working on the graph-based pan-reference will be required. However, for plant species of high within-species sequence diversity, the challenge is how to anchor large numbers of NRR sequences, whose sizes may be as large as half of the reference genome. Finally, considering the prediction of "new" or novel genes based on simple thresholds of sequence homology without detailed information on gene functionality is always somewhat arbitrary, the pangenome results based on the NGS technology can be validated and improved greatly if high-quality reference genomes of relatively few representative individuals are included in a pangenome study, particularly for important model species of relatively small genome sizes.

References

Baier U, Beller T, Ohlebusch E (2016) Graphical pan-genome analysis with compressed suffix trees and the Burrows-Wheeler transform. Bioinformatics 32:497–504

Bickhart DM, Liu GE (2014) The challenges and importance of structural variation detection in livestock. Front Genet 5:37

Bush SJ, Castillo-Morales A, Tovar-Corona JM, Chen L, Kover PX, Urrutia AO (2013) Presence–absence variation in *A. thaliana* is primarily associated with genomic signatures consistent with relaxed selective constraints. Mol Biol Evol 31:59–69

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C et al (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43:956–963

Chen W-H, Trachana K, Lercher MJ, Bork P (2012) Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol Biol Evol 29:1703–1706

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen H-C, Agarwala R, McLaren WM, Ritchie GR (2011) Modernizing reference genome assemblies. PLoS Biol 9: e1001091

Darracq A, Vitte C, Nicolas S, Duarte J, Pichon JP, Mary-Huard T, Chevalier C, Berard A, Le Paslier MC, Rogowsky P et al (2018) Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. BMC Genomics 19:119

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97

Frankel O, Brown A (1984) Current plant genetic resources – a critical appraisal. In: Chopra VL et al (eds) Genetics: new frontiers: proceedings of the XV international congress of genetics. Oxford & IBH Publishing Co., c1984, New Delhi

Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526:68–74

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CK, Severn-Ellis A, McCombie WR, Parkin IA et al (2016) The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun 7:13390

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Penagaricano F, Lindquist E, Pedraza MA, Barry K et al (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26:121–135

Hu Z, Sun C, Lu KC, Chu X, Zhao Y, Lu J, Shi J, Wei C (2017) EUPAN enables pan-genome studies of a large number of eukaryotic genomes. Bioinformatics 33:2408–2409

Hu Z, Wang W, Wu Z, Sun C, Li M, Lu J, Fu B, Shi J, Xu J, Ruan J et al (2018) Novel sequences, structural variations and gene presence variations of Asian cultivated rice. Sci Data 5:180079

James BT, Luczak BB, Girgis HZ (2018) MeShClust: an intelligent tool for clustering DNA sequences. Nucleic Acids Res 46(14):e83

Jia J, Li H, Zhang X, Li Z, Qiu L (2017) Genomics-based plant germplasm research (GPGR). Crop J 5:166–174

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J et al (2010) Building the sequence map of the human pan-genome. Nat Biotechnol 28:57–63

Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L et al (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32:1045–1052

Liu XH, Lu TT, Yu SL, Li Y, Huang YC, Huang T, Zhang L, Zhu JJ, Zhao Q, Fan DL et al (2007) A collection of 10,096 indica rice full-length cDNAs reveals highly expressed sequence divergence between *Oryza sativa indica* and *japonica* subspecies. Plant Mol Biol 65:403–415

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X et al (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun 6:6914

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101:12404–12410

Marschall T, Marz M, Abeel T, Dijkstra L, Dutilh BE, Ghaffaari A, Kersey P, Kloosterman WP, Makinen V, Novak AM et al (2018) Computational pan-genomics: status, promises and challenges. Brief Bioinform 19:118–135

Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan CK, Visendi P, Lai K, Dolezel J, Batley J, Edwards D (2017) The pangenome of hexaploid bread wheat. Plant J 90:1007–1013

Nguyen N, Hickey G, Zerbino DR, Raney B, Earl D, Armstrong J, Kent WJ, Haussler D, Paten B (2015) Building a pan-genome reference for a population. J Comput Biol 22:387–401

Ou L, Li D, Lv J, Chen W, Zhang Z, Li X, Yang B, Zhou S, Yang S, Li W (2018) Pan-genome of cultivated pepper (Capsicum) and its use in gene presence-absence variation analyses. New Phytol 220:360

Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC, Zaina G, Bastien C, Cattonaro F, Marroni F, Morgante M (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. Mol Biol Evol 33:2706–2719

Potato Genome Sequencing C, Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R et al (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189–195

Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier A, Salamov A et al (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 499:209–213

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13:278–289

Saxena RK, Edwards D, Varshney RK (2014) Structural variations in plant genomes. Brief Funct Genomics 13:296–307

Schatz MC, Maron LG, Stein JC, Hernandez Wences A, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E et al (2014) Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. Genome Biol 15:506

Schneider GF, Dekker C (2012) DNA sequencing with nanopores. Nat Biotechnol 30:326

Steinegger M, Söding J (2018) Clustering huge protein sequence sets in linear time. Nat Commun 9:2542

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH et al (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81

Sun C, Hu Z, Zheng T, Lu K, Zhao Y, Wang W, Shi J, Wang C, Lu J, Zhang D et al (2017) RPAN: rice pan-genome browser for approximately 3000 rice genomes. Nucleic Acids Res 45:597–605

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20:1689–1699

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102:13950–13955

Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F et al (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49

Yao W, Li G, Zhao H, Wang G, Lian X, Xie W (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol 16:187

Zekic T, Holley G, Stoye J (2018) Pan-genome storage and analysis techniques. Methods Mol Biol 1704:29–53

Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T et al (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50:278–284

Zheng LY, Guo XS, He B, Sun LJ, Peng Y, Dong SS, Liu TF, Jiang SY, Ramachandran S, Liu CM, Jing HC (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). Genome Biol 12:R114