Nanopore Sequencing for Infectious Disease Applications

by

Yunfan Fan

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

September, 2022

Abstract

While next generation sequencing (NGS) has enabled massively parallel DNA sequencing for lower and lower cost, the development of third generation nanopore sequencing offers several key advantages over older sequencing methods. Nanopore sequencers are pocket-sized, making them orders of magnitude cheaper than the next most affordable alternative and the ideal option for wide deployment. They are capable of providing data in real-time, saving valuable hours before data analysis can begin. Additionally, they are able to sequence reads several thousand basepairs long, as opposed to the hundreds of basepairs NGS platforms are capable of, and they embed base modification data without the need for specific treatment beforehand. Given these advantages, in this thesis I examine the application of nanopore sequencing to the study of human pathogens.

First, we use nanopore sequencing to characterize antimicrobial resistance (AMR) in forty clinical isolates. We analyzed real-time data to quickly identify AMR genes, assembled genomes to identify chromosomal mutations, and used short-read sequencing data to correct the errors in the assemblies. With sequencing data, we found that time to effective antibiotic therapy could be shortened by as much as 20 hours compared to standard antimicrobial

susceptibility testing (AST).

Second, we leverage the long reads of nanopore sequencing to assemble the genome of a pathogenic yeast, *Candida nivariensis*. Previous efforts to assemble this yeast genome relied solely on short-read NGS data, resulting in a highly fragmented genome. Using nanopore data, we achieve a much higher contiguity and capture previously missing portions of the genome. Furthermore, we demonstrate that our more contiguous genome can be used to better study long and repetative genes, such as those involved in pathogenticity to humans.

Third, we use the base modification information embedded in nanopore sequencing data to call methylation in metagenomic assemblies. These calls enable the binning of metagenomic contigs according to methylation signature without the need to collect additional data. We demonstrate the efficacy of this method on a synthetic community sample, a simple two-bacteria system, and a clinical sample with matched proximity ligation binning data.

These applications of nanopore sequencing demonstrate its potential and its utility for all fronts of pathogen genomics research.

Thesis Committee

Dr. Winston Timp (Advisor, Reader)
Associate Professor
Department of Biomedical Engineering
Department of Molecular Biology and Genetics
Johns Hopkins University School of Medicine

Dr. Patricia Simner (Reader)
Associate Professor
Department of Pathology
Johns Hopkins University School of Medicine

Dr. Steven Salzberg
Bloomberg Distinguished Professor
Department of Computer Science
Johns Hopkins University Whiting School of Engineering
Department of Biomedical Engineering
Johns Hopkins University School of Medicine
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

I have tremendous gratitude
to those people,
numerous and innumerable,
who have contributed,
directly or in subtler ways,
to this work.

Some of them are listed here.

To my advisor, Winston: I remember writing to you as a sophomore in college many years ago, asking to do research in your brand new lab, which at the time was but a few months old. Back then, I had no idea what it was to do research, and I had no relevant skills or credentials to offer, only my time and my interest to learn. Over these years I have indeed learned a lot, and I will always be grateful to you for building the place where I was able to grow.

To my thesis committee, Dr. Trish Simner and Dr. Steven Salzberg: Thank you for your exceptionally kind guidance, support, and feedback. I always left committee meetings with you feeling more confident in myself, and more optimistic about my progress.

To the @yfan arc of the #core channel - @isac, @brochael, @shao, @gilfunk, @narley, @broham, @gmoney, @Brittany, @sherbear, @Sam Sholes, @Paul Hook, @amymeltzer39, @alice, @Luke Morina, @Courtney Johnson, and @Jess: Thank you for those times when you patiently watched over me as I learned new lab techniques, answered my questions, and rescued me from predicaments of my own making. It is my fondest hope that at some point during our time together, I was able to be mildly helpful to you as well. Thank you most of all for commiserating with me as we struggled together through the singular challenges of grad school, and celebrating the equally singular triumphs.

To the crew that moved me into 703 (and Charles, and Charlotte, and Sven, and Manolo, and the tumbledonkses that I've loved and lost): Thanks for being there, and thanks for hanging out. Let's go climbing and get a beer the next time we're all around. It's been a while.

To mom and dad, and family further away: It was your labor that first cultivated my growth. Accomplishments in my name are as much yours as they are mine. I flourish for you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Sequencing Technology

Since Sanger developed a chain-terminating procedure for DNA sequencing over forty years ago (Sanger, Nicklen, and Coulson, 1977), sequencing capabilities have grown, at first steadily, then astronomically (Schatz and Langmead, 2013). The advent of sequencing-by-synthesis ushered in 'next-generation' sequencing (NGS) methods, whereby DNA sequencing became 'massively parallel' in nature and vastly accessible to researchers in all fields of biology and medicine. These NGS methods typically involve immobilizing millions of DNA fragments, amplifying them, and then observing the activity of DNA polymerase as it synthesizes the complement strands to the amplified fragments. Commonly, this observation is done through imaging fluorescently labeled nucleotides one base addition, or cycle, at a time (Shendure et al., 2017).

The highly democratized nature of NGS has enabled researchers in clinical microbiology to use it for a variety of important applications. These range from

cataloging and surveilling genetic determinants of antimicrobial resistance (AMR) (Crofts, Gasparrini, and Dantas, 2017; Caniça et al., 2019; Tóth et al., 2020; Thanner, Drissner, and Walsh, 2016; Hendriksen et al., 2019), to monitoring outbreaks of infectious diseases (Di Paola et al., 2020; Lu et al., 2020), to analyzing entire human microbiomes with metagenomic sequencing (Chiu and Miller, 2019). Based on the advances brought about by NGS, some have even called for establishing a 'digital immune system' whereby sequencing-based microbial surveillance would detect threats of outbreak, which then could be contained before they become too difficult to control (Schatz and Phillippy, 2012).

While NGS technology has unlocked enormous advances in clinical microbiology, it is limited by the fragment lengths it can handle and its reliance on amplification. Typical NGS methods are unreliable at sequencing individual DNA fragments multiple thousands of base pairs long (Heather and Chain, 2016), and can only read stretches of a few hundred nucleotides at a time. These short read lengths make the sequencing data difficult to work with for many downstream applications, including genome assembly and analysis of repetitive genomic loci. Meanwhile, the amplification process obliterates any base modification information, such as methylation, potentially present on the native DNA fragment.

The rise of third generation, single-molecule sequencing just in the past decade has begun to address these shortcomings. These methods, also known as 'long-read' sequencing, interrogate individual DNA molecules without the need for amplification, and can read stretches of thousands of nucleotides

at a time (Jain et al., 2018). Nanopore sequencing in particular does this by measuring the minute fluctuations in ionic current as a single stranded DNA molecule passes through a transmembrane protein pore. Because no amplification or other chemical treatments are required prior to sequencing, base modification information remains intact and can be read simultaneously with the nucleotide sequence (Simpson et al., 2017; McIntyre et al., 2019). Also unlike NGS methods, the current single molecule sequencing procedures are not dependent on imaging single bases from all the reads at once in a synchronized fashion. Data is collected from all pores independently, which enables data streaming to analysis pipelines even as more data are still being collected. To further the role of sequencing for infectious disease applications, I leverage the new capabilities of nanopore sequencing for AMR detection, genome assembly of eukaryotic pathogens, and metagenomics.

1.2 Antimicrobial resistance

Since Alexander Fleming first observed the 'bacteriolytic' properties of a mysterious 'mould broth filtrate' which he termed 'penicillin' (Fleming, 1929), antibiotics have been an unprecedented and miraculous silver bullet against previously deadly bacteria. Usually produced and isolated from fungi (Martínez, 2008), antibiotics are small molecules capable of killing bacteria or inhibiting their growth without damaging eukaryotic cells or tissues in the vicinity. Not only are antibiotics used to cure infectious diseases, but they have enabled more and more complex medical interventions such as surgery and chemotherapy by drastically reducing the risk and ramification of bacterial infections

(Crofts, Gasparrini, and Dantas, 2017). Outside of medicine, antimicrobials have also been used extensively in animal agriculture to control disease (Aarestrup, 2015), as populations of food animals are scaled to accommodate global diets that continue to demand more animal protein (Van Boeckel et al., 2019). Consequently, antibiotics are one of the most commonly prescribed classes of drugs in recent decades, and their use is only becoming more widespread (Van Boeckel et al., 2014).

However, for as long as fungi have produced antimicrobials, bacteria have evolved resistances to them (D'Costa et al., 2011). As human usage of antibiotics has occurred ubiquitously and without restraint, selection pressures have caused a rapid proliferation of antibiotic resistant strains of bacteria. Even synthetic antibiotics such as quinolones sustained only three decades of widespread usage before resistances began to develop, intensify, and spread (Laxminarayan et al., 2013; Ruiz, Pons, and Gomes, 2012). Multidrug-resistant strains of bacteria have also emerged and become prominent, deepening the international crisis of AMR (Tamma and Cosgrove, 2014).

In the clinic, the prevalence of AMR makes it difficult to immediately prescribe the most effective interventions for patients colonized with commonly drug resistant bacteria. Not only does this cost potentially crucial time from patient treatment, but it could cause antibiotic waste and contribute to selection pressures causing drug resistances to develop in the first place. In Chapter 2, I explore the use of third generation sequencing in the clinic to rapidly detect drug resistances in order to shorten the time to effective antibiotic therapy and enable antibiotic stewardship.

1.3 Genome Assembly

High quality, complete genomes are crucial not only for population and comparative genomics, but they also commonly underpin gene expression studies, epigenetics assays, and molecular diagnostics (Rhie et al., 2021). Because highly parallel sequencing technologies cannot record whole genomes on a single read, genomes must be reconstructed out of millions or billions of reads. This process has been likened to a large jigsaw puzzle, where reads must be overlapped, oriented and fit together in order to build the larger picture of the genome (Sohn and Nam, 2018). Contiguous sequences constructed by overlapping reads in this fashion are known as 'contigs,' which typically represent large sections of chromosomes.

Repetitive and low complexity regions of the genome have been difficult to resolve using NGS technologies. The short reads often cannot span these regions, making it difficult to unambiguously determine how long they are, and how many repeats each region contains (Paszkiewicz and Studholme, 2010). Contigs are typically terminated at these ambiguous regions, resulting in highly fragmented assemblies containing thousands of contigs. Long read data capable of spanning repetitive regions are able to resolve the ambiguities they cause, resulting in much more contiguous genome assemblies with longer and fewer contigs. Genome assemblies constructed from only long read data are more prone to single-base errors due to the lower accuracy of the long reads, but NGS data gathered on the same sample can be used to correct most small errors (Goodwin et al., 2015). Using both long-read and NGS data leverages the benefits and addresses the weaknesses of both sequencing

technologies.

Only with contiguous, reference-quality genome assemblies can the roles of repeat structures and long, repetitive genes be analyzed. In pathogenic fungi, some adhesion proteins tend to be encoded in long genes with tandem repeats embedded within. These proteins are of particular interest as they are thought to be involved in enabling pathogenicity (Timmermans et al., 2018). In Chapter 3, I use both NGS and long-read data to assemble the genome of *Candida nivariensis*, a pathogenic yeast, and use the assembled genome to explore the long, repetitive genes encoding adhesions in this species.

1.4 Metagenomics

Microbes are ubiquitous, and structured microbial communities, or microbiomes, can be found associating with a variety of hosts and environmental niches (Quince et al., 2017). Increasingly, the human associated microbiomes are found to play an important role in human health (Fan and Pedersen, 2021). Studying complex communities using traditional microbial methods based on culturing bacteria has been difficult, as not all microbes can be cultured, and the culturing process itself would be likely to alter the composition of the community (Quince et al., 2017).

Community members can be identified and quantified using 16S rRNA sequencing, whereby the 16S genes of all microbial genomes in the community are simultaneously amplified, and then sequenced. By comparing only these 16S sequences to each other and to reference databases, operational taxonomic units (OTUs) making up the community can be determined, and taxonomy

can be assigned (Johnson et al., 2019). While 16S-based methods are effective for studying the organism composition of microbiomes, it cannot directly shed any light on the functional capabilities of the microbes.

By contrast, metagenomics sequencing captures the full genetic complement of the community, including any genes, plasmids, or phages which may be present in the cells and their environs. While more sequences are captured with metagenomics, analyzing this data becomes more difficult computationally (Breitwieser, Lu, and Salzberg, 2019). One common approach to analysis involves assembling the metagenome in order to determine identity and functions of the microbes in the community (Lapidus and Korobeynikov, 2021). Metagenome assembly is functionally similar to genome assembly of a single organism, where reads are overlapped in order to construct contigs. However, because metagenomic contigs can originate from an undetermined number of organisms, grouping, or 'binning,' these contigs according to the species of origin is a crucial yet challenging step in metagenomic analysis (Yue et al., 2020).

Binning metagenomic contigs into metagenome assembled genomes (MAGs), using NGS data has typically been done using a combination of the contigs' kmer-spectra, and differential coverage (Ghurye, Cepeda-Espinoza, and Pop, 2016). While these methods work well for bacterial chromosomes, they are less effective for binning mobile genetic elements (MGEs), especially if these MGEs are capable of replicating independently of the host chromosome, as most plasmids are. As with genome assembly of a single organism, the use of NGS data in metagenomic assembly limits the lengths of the contigs themselves,

potentially resulting in unresolvable repeats and truncated gene sequences.

By applying long-read sequencing to metagenomic analysis, it is possible to assemble much longer contigs from microbiome samples. Furthermore, because native base modification information is preserved, it can be a powerful basis for binning, as modifications are preserved on MGEs and are not affected by chromosome-independent replication. In Chapter 4, I use methylation calls derived from nanopore sequencing for metagenomic binning, and assess its effectiveness.

References

Sanger, F, S Nicklen, and A R Coulson (1977). "DNA sequencing with chain-terminating inhibitors". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 74.12, pp. 5463–5467.

Schatz, Michael C and Ben Langmead (2013). "The DNA Data Deluge: Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze". en. In: *IEEE Spectrum* 50.7, pp. 26–33.

Shendure, Jay, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston (2017). "DNA sequencing at 40: past, present and future". en. In: *Nature* 550.7676, pp. 345–353.

Crofts, Terence S, Andrew J Gasparrini, and Gautam Dantas (2017). "Next-generation approaches to understand and combat the antibiotic resistome". en. In: *Nat. Rev. Microbiol.* 15.7, pp. 422–434.

Caniça, Manuela, Vera Manageiro, Hikmate Abriouel, Jacob Moran-Gilad, and Charles M A P Franz (2019). "Antibiotic resistance in foodborne bacteria". In: *Trends Food Sci. Technol.* 84, pp. 41–44.

Tóth, Adrienn Gréta, István Csabai, Eszter Krikó, Dóra Tőzsér, Gergely Maróti, Árpád V Patai, László Makrai, Géza Szita, and Norbert Solymosi (2020). "Antimicrobial resistance genes in raw milk for human consumption". en. In: *Sci. Rep.* 10.1, p. 7464.

Thanner, Sophie, David Drissner, and Fiona Walsh (2016). "Antimicrobial Resistance in Agriculture". en. In: *MBio* 7.2, e02227–15.

Hendriksen, Rene S, Patrick Munk, Patrick Njage, Bram van Bunnik, Luke McNally, Oksana Lukjancenko, Timo Röder, David Nieuwenhuijse, Susanne Karlsmose Pedersen, Jette Kjeldgaard, Rolf S Kaas, Philip Thomas Lanken Conradsen Clausen, Josef Korbinian Vogt, Pimlapas Leekitcharoenphon, Milou G M van de Schans, Tina Zuidema, Ana Maria de Roda Husman, Simon Rasmussen, Bent Petersen, Global Sewage Surveillance project consortium, Clara Amid, Guy Cochrane, Thomas Sicheritz-Ponten, Heike

Schmitt, Jorge Raul Matheu Alvarez, Awa Aidara-Kane, Sünje J Pamp, Ole Lund, Tine Hald, Mark Woolhouse, Marion P Koopmans, Håkan Vigre, Thomas Nordahl Petersen, and Frank M Aarestrup (2019). "Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage". en. In: *Nat. Commun.* 10.1, p. 1124.

Di Paola, Nicholas, Mariano Sanchez-Lockhart, Xiankun Zeng, Jens H Kuhn, and Gustavo Palacios (2020). "Viral genomics in Ebola virus research". en. In: *Nat. Rev. Microbiol.* 18.7, pp. 365–378.

Lu, Jing, Louis du Plessis, Zhe Liu, Verity Hill, Min Kang, Huifang Lin, Jiufeng Sun, Sarah François, Moritz U G Kraemer, Nuno R Faria, John T McCrone, Jinju Peng, Qianling Xiong, Runyu Yuan, Lilian Zeng, Pingping Zhou, Chumin Liang, Lina Yi, Jun Liu, Jianpeng Xiao, Jianxiong Hu, Tao Liu, Wenjun Ma, Wei Li, Juan Su, Huanying Zheng, Bo Peng, Shisong Fang, Wenzhe Su, Kuibiao Li, Ruilin Sun, Ru Bai, Xi Tang, Minfeng Liang, Josh Quick, Tie Song, Andrew Rambaut, Nick Loman, Jayna Raghwani, Oliver G Pybus, and Changwen Ke (2020). "Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China". en. In: *Cell* 181.5, 997–1003.e9.

Chiu, Charles Y and Steven A Miller (2019). "Clinical metagenomics". en. In: *Nat. Rev. Genet.* 20.6, pp. 341–355.

Schatz, Michael C and Adam M Phillippy (2012). "The rise of a digital immune system". en. In: *Gigascience* 1.1, p. 4.

Heather, James M and Benjamin Chain (2016). "The sequence of sequencers: The history of sequencing DNA". en. In: *Genomics* 107.1, pp. 1–8.

Jain, Miten, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads". en. In: *Nat. Biotechnol.* 36.4, pp. 338–345.

Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp (2017). "Detecting DNA cytosine methylation using nanopore sequencing". en. In: *Nat. Methods* 14.4, pp. 407–410.

McIntyre, Alexa BR, Noah Alexander, Kirill Grigorev, Daniela Bezdan, Heike Sichtig, Charles Y Chiu, and Christopher E Mason (2019). "Single-molecule sequencing detection of N6-methyladenine in microbial reference materials". In: *Nature communications* 10.1, pp. 1–11.

Fleming, Alexander (1929). "On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of *B. influenzæ*". en. In: *Br. J. Exp. Pathol.* 10.3, p. 226.

Martínez, José L (2008). "Antibiotics and antibiotic resistance genes in natural environments". en. In: *Science* 321.5887, pp. 365–367.

Aarestrup, Frank M (2015). "The livestock reservoir for antimicrobial resistance: a personal view on changing patterns of risks, effects of interventions and the way forward". en. In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370.1670, p. 20140085.

Van Boeckel, Thomas P, João Pires, Reshma Silvester, Cheng Zhao, Julia Song, Nicola G Criscuolo, Marius Gilbert, Sebastian Bonhoeffer, and Ramanan Laxminarayan (2019). "Global trends in antimicrobial resistance in animals in low- and middle-income countries". en. In: *Science* 365.6459.

Van Boeckel, Thomas P, Sumanth Gandra, Ashvin Ashok, Quentin Caudron, Bryan T Grenfell, Simon A Levin, and Ramanan Laxminarayan (2014). "Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data". en. In: *Lancet Infect. Dis.* 14.8, pp. 742–750.

D'Costa, Vanessa M, Christine E King, Lindsay Kalan, Mariya Morar, Wilson W L Sung, Carsten Schwarz, Duane Froese, Grant Zazula, Fabrice Calmels, Regis Debruyne, G Brian Golding, Hendrik N Poinar, and Gerard D Wright (2011). "Antibiotic resistance is ancient". en. In: *Nature* 477.7365, pp. 457–461.

Laxminarayan, Ramanan, Adriano Duse, Chand Wattal, Anita K M Zaidi, Heiman F L Wertheim, Nithima Sumpradit, Erika Vlieghe, Gabriel Levy Hara, Ian M Gould, Herman Goossens, Christina Greko, Anthony D So, Maryam Bigdeli, Göran Tomson, Will Woodhouse, Eva Ombaka, Arturo Quizhpe Peralta, Farah Naz Qamar, Fatima Mir, Sam Kariuki, Zulfiqar A Bhutta, Anthony Coates, Richard Bergstrom, Gerard D Wright, Eric D Brown, and Otto Cars (2013). "Antibiotic resistance-the need for global solutions". en. In: *Lancet Infect. Dis.* 13.12, pp. 1057–1098.

Ruiz, Joaquim, Maria J Pons, and Cláudia Gomes (2012). "Transferable mechanisms of quinolone resistance". en. In: *Int. J. Antimicrob. Agents* 40.3, pp. 196–203.

Tamma, Pranita D and Sara E Cosgrove (2014). "Let the games begin: the race to optimise antibiotic use". en. In: *Lancet Infect. Dis.* 14.8, pp. 667–668.

Rhie, Arang et al. (2021). "Towards complete and error-free genome assemblies of all vertebrate species". en. In: *Nature* 592.7856, pp. 737–746.

Sohn, Jang-Il and Jin-Wu Nam (2018). "The present and future of de novo whole-genome assembly". en. In: *Brief. Bioinform.* 19.1, pp. 23–40.

Paszkiewicz, Konrad and David J Studholme (2010). "De novo assembly of short sequence reads". en. In: *Brief. Bioinform.* 11.5, pp. 457–472.

Goodwin, Sara, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C Schatz, and W Richard McCombie (2015). "Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome". en. In: *Genome Res.* 25.11, pp. 1750–1756.

Timmermans, Bea, Alejandro De Las Peñas, Irene Castaño, and Patrick Van Dijck (2018). "Adhesins in *Candida glabrata*". en. In: *J Fungi (Basel)* 4.2.

Quince, Christopher, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata (2017). "Shotgun metagenomics, from sampling to analysis". en. In: *Nat. Biotechnol.* 35.9, pp. 833–844.

Fan, Yong and Oluf Pedersen (2021). "Gut microbiota in human metabolic health and disease". en. In: *Nat. Rev. Microbiol.* 19.1, pp. 55–71.

Johnson, Jethro S, Daniel J Spakowicz, Bo-Young Hong, Lauren M Petersen, Patrick Demkowicz, Lei Chen, Shana R Leopold, Blake M Hanson, Hanako O Agresta, Mark Gerstein, Erica Sodergren, and George M Weinstock (2019). "Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis". en. In: *Nat. Commun.* 10.1, p. 5029.

Breitwieser, Florian P, Jennifer Lu, and Steven L Salzberg (2019). "A review of methods and databases for metagenomic classification and assembly". en. In: *Brief. Bioinform.* 20.4, pp. 1125–1136.

Lapidus, Alla L and Anton I Korobeynikov (2021). "Metagenomic Data Assembly - The Way of Decoding Unknown Microorganisms". en. In: *Front. Microbiol.* 12, p. 613791.

Yue, Yi, Hao Huang, Zhao Qi, Hui-Min Dou, Xin-Yi Liu, Tian-Fei Han, Yue Chen, Xiang-Jun Song, You-Hua Zhang, and Jian Tu (2020). "Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets". en. In: *BMC Bioinformatics* 21.1, p. 334.

Ghurye, Jay S, Victoria Cepeda-Espinoza, and Mihai Pop (2016). "Metagenomic Assembly: Overview, Challenges and Applications". en. In: *Yale J. Biol. Med.* 89.3, pp. 353–362.

Chapter 2

Applying Rapid Whole-Genome Sequencing To Predict Phenotypic Antimicrobial Susceptibility Testing Results among Carbapenem-Resistant *Klebsiella pneumoniae* Clinical Isolates

Portions of this chapter originally appeared in:

Tamma PD, Fan Y, Bergman Y, Pertea G, Kazmi AQ, Lewis S, et al. Applying Rapid Whole-Genome Sequencing To Predict Phenotypic Antimicrobial Susceptibility Testing Results among Carbapenem-Resistant *Klebsiella pneumoniae* Clinical Isolates 2019;63. https://doi.org/10.1128/AAC.01923-18

2.1 Abstract

Standard antimicrobial susceptibility testing (AST) approaches lead to delays in the selection of optimal antimicrobial therapy. Here, we sought to determine the accuracy of antimicrobial resistance (AMR) determinants identified by Nanopore whole-genome sequencing in predicting AST results. Using a cohort of 40 clinical isolates (21 carbapenemase-producing carbapenem-resistant *Klebsiella pneumoniae*, 10 non-carbapenemase-producing carbapenem-resistant *K. pneumoniae*, and 9 carbapenem-susceptible *K. pneumoniae* isolates), three separate sequencing and analysis pipelines were performed, as follows: (i) a real-time Nanopore analysis approach identifying acquired AMR genes, (ii) an assembly-based Nanopore approach identifying acquired AMR genes and chromosomal mutations, and (iii) an approach using short-read correction of Nanopore assemblies. The short-read correction of Nanopore assemblies served as the reference standard to determine the accuracy of Nanopore sequencing results. With the real-time analysis approach, full annotation of acquired AMR genes occurred within 8h from subcultured isolates. Assemblies sufficient for full resistance gene and single-nucleotide polymorphism annotation were available within 14h from subcultured isolates. The overall agreement of genotypic results and anticipated AST results for the 40 *K. pneumoniae* isolates was 77% (range, 30% to 100%) and 92% (range, 80% to 100%) for the real-time approach and the assembly approach, respectively. Evaluating the patients contributing the 40 isolates, the real-time approach and assembly approach could shorten the median time to effective antibiotic therapy by 20h and 26h, respectively, compared to standard AST. Nanopore

sequencing offers a rapid approach to both accurately identify resistance mechanisms and to predict AST results for *K. pneumoniae* isolates. Bioinformatics improvements enabling real-time alignment, coupled with rapid extraction and library preparation, will further enhance the accuracy and workflow of the Nanopore real-time approach.

2.2 Introduction

Whole-genome sequencing (WGS) has enabled notable advancements to the field of infectious diseases, such as an improved understanding of transmission dynamics and outbreak analysis (Didelot et al., 2012). An exciting possibility from this technology is the ability to predict antimicrobial susceptibility testing (AST) results based on the identification of acquired resistance genes and/or chromosomal mutations (Shelburne et al., 2017).

Currently, there are several shortcomings with standard approaches to AST, particularly as they relate to multidrug-resistant Gram-negative (MDRGN) organisms. First, AST results are reported approximately 48 to 72h after the time of culture collection, potentially leading to delays in appropriate empirical antibiotic therapy (Caliendo et al., 2013). Second, automated AST panels are limited in the number of antibiotic agents included. For agents that frequently need to be considered for highly drug-resistant pathogens (e.g., colistin, tigecycline, ceftazidime-avibactam, etc.) and newer agents in later phases of development that are unlikely to be routinely included in AST panels for the foreseeable future, there are additional delays in AST determination. As it is generally not evident at the time antibiotics are initiated that a patient will be

infected with an MDRGN organism, susceptibility testing for these last-resort agents occurs subsequent to, and not simultaneously with, automated AST testing. Third, standard AST reporting does not include identification of resistance mechanisms (e.g., carbapenemases, extended-spectrum β-lactamases [ESBLs], etc.), which can be important for guiding antibiotic treatment decisions, as *in vitro* activity does not always translate to *in vivo* activity (Tamma et al., 2017). WGS can potentially alleviate many of these concerns by offering the potential to predict AST results by identifying the presence or absence of resistance genes, as well as mutations in relevant genes, from which clinicians can infer the activity of antibiotic agents. Furthermore, once sequencing data has been collected, it can be used to place bacterial isolates in the context of previously acquired data, which can be useful for genomic epidemiological studies and surveillance.

Oxford Nanopore Technologies (Oxford, England) has created a Nanopore-based DNA sequencer that sequences DNA by monitoring the electrical current as DNA passes through a protein pore. Unlike second-generation sequencing methods, which require the entire run to be completed before data can be analyzed, Nanopore sequencing streams long-read data in real time (Cao et al., 2016), allowing for resistance gene identification within as few as 15min of beginning the sequencing run (Schmidt et al., 2017; Lemon et al., 2017; Judge et al., 2015). As the duration of time needed for DNA extraction and library preparation techniques continues to be reduced, the total time to identification of resistance determinants from organism growth could conceivably be accomplished within a single laboratory shift. To further advance

this science, we evaluated the correlation of resistance determinants identified through Nanopore sequencing with AST results in a cohort of 40 clinical *Klebsiella pneumoniae* complex isolates (**Figure** 2.1). This also enabled us to quantify the potential decrease in time to effective antibiotic therapy for the patients contributing isolates with the use of WGS using real-time analysis or rapid assembly approaches compared to that with traditional AST methods.

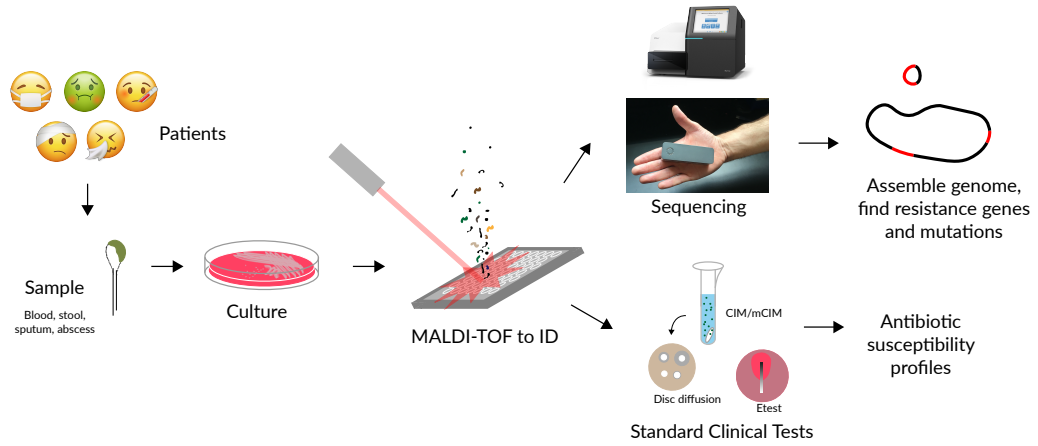

Figure 2.1: Study overview. Samples were collected from patients, and bacterial isolates were cultured and identified. Isolates were then sequenced in parallel with undergoing phenotypic AST.

As AMR can result from plasmid-mediated gene acquisition or point mutations to drug targets (**Figure** 2.2), detecting both is vital to predicting resistance phenotypes from genotypic data. While real-time analysis of nanopore data can enable the detection of whole genes within minutes, individual reads remain too error prone to detect small point mutations. In order to detect these, consensus sequences such as those generated by genome assembly must be used (**Figure** 2.3). While this vastly improves sequence accuracy, most systematic errors around homopolymers and methylation motifs have been found

to persist (**Figure** 2.4). To address this, we further corrected the assemblies using nanopolish, which leverages the raw electrical signal produced by the sequencer to improve consensus accuracy.

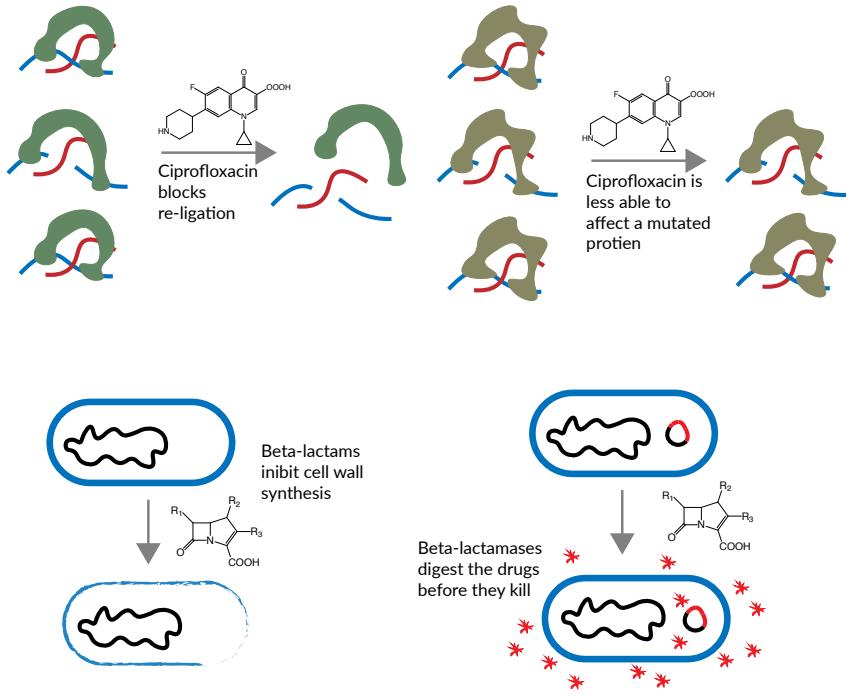

Figure 2.2: Resistance mechanisms. Drug resistances can arise through acquisition of a protective gene, such a β-lactamase gene. Mutations to a drug target such as the *gyrA* topoisomerase gene can also reduce the effectiveness of antimicrobial agents.

2.3 Results

2.3.1 Sequencing runs and genome assemblies

The nanopore sequencing runs yielded a mean of 6.4Gbp of sequencing data (**Table** 2.1), with all but one yielding at least 1 Gbp. As the *Klebsiella pneumoniae* genome is only 5Mbp, this mean yield accounts for a greater than 1000X

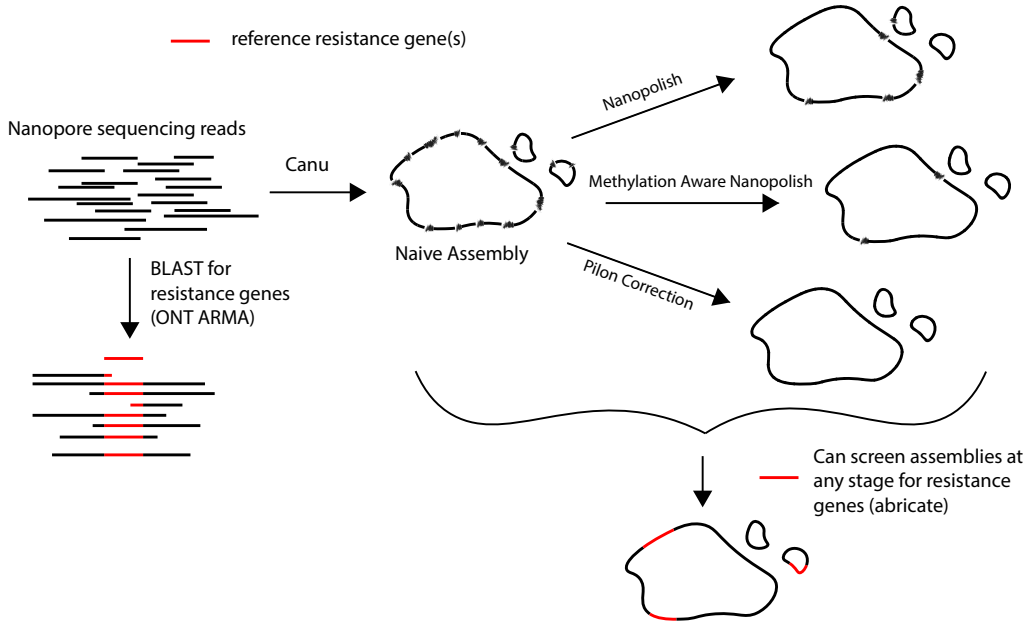

Figure 2.3: Sequencing analysis pipeline. Sequencing reads can be searched in real-time for resistance genes. To detect and assess point mutations, genomes are assembled, and then can be polished using native electrical data (Nanopolish), or highly accurate short-read data (Pilon).

sequence coverage per isolate. Illumina sequencing runs yielded an average of 201Mbp per run (**Table** 2.2), resulting in an average of 40X coverage. Using the regular assembly pipeline, 38 of the 40 genome assemblies appear to contain a full length bacterial chromosome at least 5Mbp long (**Figure** 2.5**, Table** 2.3). Further polishing steps both with and without Illumina data do not significantly affect the contig lengths in the assemblies (**Tables** 2.4**,** 2.5). With the rapid analysis pipeline using downsampled data, 35 of the 40 assemblies contain a chromosome length contig (**Table** 2.6). Using the illumina-corrected assemblies, we were able to place our isolates in the context of other previously published, chromosome-level assemblies available in GenBank (**Figure** 2.6).

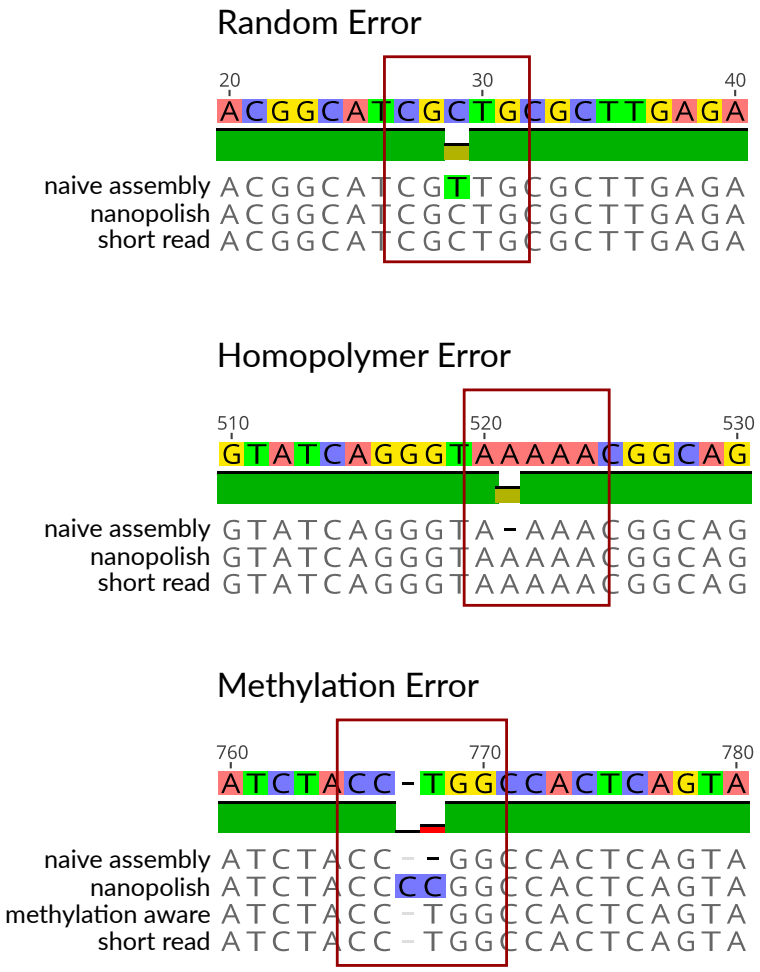

Figure 2.4: Correction examples. Nanopolish corrects most random errors which persist after genome assembly to agree with short-read correction. It also corrects homopolymer errors, the most prominent systematic error on the ONT platform. Methylation-associated errors, such as those in the context of the dcm methyltransferase (CCWGG), are more effectively corrected using the methylation-aware mode.

2.3.2 Percent agreement of WGS in predicting AST results

The overall agreement of genotypic results and anticipated AST results for the 40 *K. pneumoniae* isolates was 77% (range, 30% to 100%), 92% (range, 80% to 100%), and 92% (range, 80% to 100%) for the Nanopore real-time approach,

Isolate Number	yield (Gbp)	# of reads	mean length	median length	mean base quality
1	8.294	1257157	6597.1	7375	7.6
2	6.116	1048031	5835.72	7480	6.84
3	0.479	77390	6187.79	7426	7.84
4	8.239	1548663	5319.84	6086	7.35
5	8.508	1217218	6989.66	7504	7.6
6	3.259	505396	6448.3	7754	6.72
7	0.164	26932	6082.12	7138	7.77
8	3.975	509581	7799.9	8504	7.67
9	8.082	1035735	7803.25	7382	8.27
10	7.724	1009608	7650.62	8334	8.05
11	8.798	1347446	6529.11	8031	6.92
12	13.805	1674043	8246.34	9293	8.49
13	1.165	140662	8278.98	10344	7.12
14	5.92	830408	7128.99	7090	8.44
15	4.197	621459	6753.05	7769	8.41
16	9.192	1246740	7373.18	7874	7.84
17	10.267	1556852	6594.51	7950	7.04
18	9.076	1134570	7999.15	9041	7.64
19	6.208	1047150	5928.37	7812	6.87
20	7.096	920706	7706.61	8725	7.65
21	11.295	1713000	6593.69	7213	7.34
22	6.363	1037059	6135.85	7462	7.22
23	5.434	750022	7244.87	8915	6.94
24	3.927	1155910	3397.65	1363	8.35
25	7.628	1132613	6735.02	7662	7.49
26	8.85	968657	9136.41	10258	8.05
27	11.858	1672521	7089.6	8479	7.92
28	4.087	922094	4431.87	4102	7.27
29	5.349	723311	7395.12	7971	8.02
30	8.583	1128360	7606.85	8350	7.84
31	5.117	650864	7861.43	9641	7
32	4.478	632788	7077.14	7927	7.25
33	5.923	813933	7277.19	9064	6.88
34	7.87	1137405	6919.49	7210	7.58
35	3.907	556051	7026.83	8079	6.9
36	4.853	831661	5835.08	6864	6.39
37	3.678	509408	7219.5	8747	6.77
38	3.848	477177	8064.5	10565	6.66
39	5.628	944401	5959.33	6340	6.72
40	6.137	977279	6279.95	6890	7.41

Table 2.1: Nanopore sequencing data. Summary of nanopore sequencing data for each isolate

the Nanopore sequencing assembly approach, and the Pilon-corrected Illumina approach, respectively (**Table** 2.7). The Nanopore real-time approach, compared to the assembly-based approach, had an inability to identify allelic variants (i.e., bla_{KPC} was identified but $bla_{\text{KPC-3}}$ could not be specifically identified). Because all *K. pneumoniae* isolates are known to have chromosomally integrated non-ESBL β-lactamase genes (e.g., $bla_{\text{SHV-1}}$, $bla_{\text{SHV-11}}$), when

Isolate Number	Yield (bp)	# of reads	mean length	median length	mean base quality
1	73189038	704430	103.9	76	35.05
2	75035866	996368	75.31	76	36.52
3	97466000	876404	111.21	76	35
4	87314517	799918	109.15	76	35.38
5	191969186	1758748	109.15	76	34.73
6	167757889	1528680	109.74	76	33.5
7	142025604	1258692	112.84	85	34.51
8	68405758	709964	96.35	76	35.74
9	51272506	448484	114.32	143	34.55
10	147679034	1395172	105.85	76	35.56
11	271473852	2294022	118.34	151	33.39
12	121803348	1092120	111.53	76	34.52
13	91725973	785506	116.77	150	33.94
14	116921911	1038468	112.59	76	34.98
15	96724518	888242	108.89	76	34.42
16	66088407	668446	98.87	76	34.83
17	88485971	775996	114.03	142	35.33
18	82855225	857310	96.65	76	35.68
19	102426035	1091590	93.83	76	35.78
20	181283343	1399544	129.53	151	35.26
21	168435381	1525782	110.39	76	33.66
22	72585638	768092	94.5	76	35.09
23	139415763	1435732	97.1	76	35.48
24	167010694	1408158	118.6	151	35.44
25	145099819	1261896	114.99	149	33.96
26	1162905519	4733052	245.7	301	30.53
27	1147954539	4477498	256.38	301	30
28	1326820865	7121026	186.32	168	31.62
29	121145118	1270030	95.39	76	35.21
30	89553255	743270	120.49	151	35.04
31	138841750	1116386	124.37	151	34.02
32	235049943	2001742	117.42	150	35.37
33	119647068	1104940	108.28	76	35.55
34	75318316	603428	124.82	151	35.34
35	139314063	1350424	103.16	76	34.19
36	53435264	496334	107.66	76	34.95
37	94411304	923498	102.23	76	34.61
38	141578108	1156864	122.38	151	35.53
39	156093332	1627956	95.88	76	34.72
40	53897028	429144	125.59	151	33.98

Table 2.2: Illumina sequencing data. Summary of Illumina sequencing data for each isolate

bla$_{\text{SHV}}$ was identified, the assumption was that it was a non-ESBL *bla*$_{\text{SHV}}$. This led to reductions in the accuracy of predictions for several β-lactams, including 5%, 2%, and 3% reductions in accurate predictions for piperacillin-tazobactam, ceftriaxone, and cefepime, respectively. Also, since the number of aminoglycoside-modifying enzymes were important to predict aminoglycoside resistance, and the alleles could not be distinguished, this led to

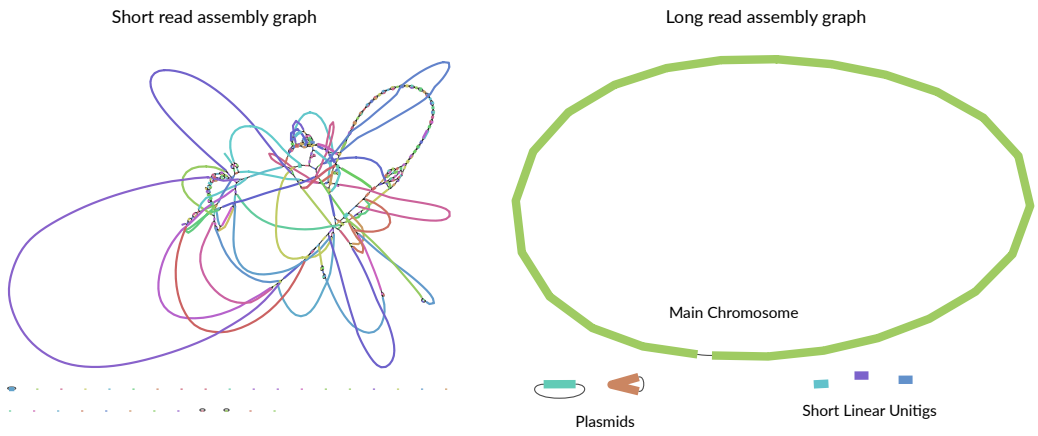

Figure 2.5: Assembly graphs. Representative assembly graphs when using only short reads and only long reads. Short reads are difficult to use for resolving repetitive regions, resulting in bubbles in the graph, whereas assemblies from long read data typically result in a full length bacterial chromosome.

decreases in accurate predictions for amikacin (reduced by 7%) and gentamicin (reduced by 48%). Additionally, the real-time approach was unable to identify chromosomal mutations, leading to decreases in accurate predictions for ciprofloxacin/levofloxacin (reduced by 68%) and colistin (reduced by 5%).

2.3.3 Time to resistance determination

With a Nanopore real-time analysis approach, acquired resistance genes were identified within 8h from subcultured isolates. Assemblies sufficient for full resistance gene and single-nucleotide polymorphism annotation using Nanopore sequences (Nanopore assembly approach) were available within 14h from subcultured isolates. **Figure** 2.7 compares the Nanopore real-time analysis and assembly approaches with standard-of-care methods. **Table** 2.7 summarizes the agreements between antibiotic resistance determinants identified using the real-time approach, assembly-based approach, or hybrid

Isolate Number	# of contigs	n50 (bp)	longest contig (bp)	shortest contig (bp)
1	4	5345918	5345918	50811
2	6	5333434	5333434	34287
3	4	5273185	5273185	8815
4	3	3985336	3985336	38998
5	2	5305648	5305648	143607
6	4	5472794	5472794	56835
7	6	5445103	5445103	18478
8	4	5257134	5257134	29283
9	5	5518371	5518371	37966
10	2	5198744	5198744	220330
11	2	5201156	5201156	202178
12	10	3270806	3270806	22754
13	5	5319848	5319848	30299
14	6	5357061	5357061	46213
15	3	5360808	5360808	73121
16	4	5486236	5486236	55582
17	4	5506760	5506760	51642
18	5	5415662	5415662	33474
19	6	5413946	5413946	23113
20	8	5297228	5297228	31483
21	8	5283730	5283730	33034
22	6	5334448	5334448	23396
23	8	5378379	5378379	41166
24	5	5338906	5338906	20744
25	1	5189390	5189390	5189390
26	3	5368726	5368726	99287
27	4	5351289	5351289	68537
28	3	5450019	5450019	212463
29	6	5406830	5406830	30987
30	4	5448768	5448768	68173
31	5	5215120	5215120	33818
32	3	5194647	5194647	134082
33	6	5257573	5257573	37028
34	4	5224706	5224706	37931
35	4	5347128	5347128	54724
36	4	5416383	5416383	35362
37	5	5392598	5392598	36539
38	3	5087129	5087129	155866
39	2	5260840	5260840	75691
40	3	5207945	5207945	94313

Table 2.3: Raw assembly statistics. Summary statistics for genome assemblies of all isolates with no further correction

Nanopore-Illumina assemblies and AST predictions.

There were 28 patients in the cohort infected with carbapenem-resistant *K. pneumoniae* strains. Overall, 22 (79%) received empirical antibiotic therapy that was not active against their infecting isolates. Results from Nanopore

Isolate Number	# of contigs	n50 (bp)	longest contig (bp)	shortest contig (bp)
1	4	5387670	5387670	51205
2	6	5379385	5379385	34533
3	4	5312070	5312070	8820
4	3	4016871	4016871	39226
5	2	5351610	5351610	144501
6	4	5517181	5517181	57263
7	6	5487393	5487393	18478
8	4	5301136	5301136	29284
9	5	5556690	5556690	37975
10	2	5241082	5241082	221896
11	2	5251331	5251331	203995
12	10	3292545	3292545	22770
13	5	5358195	5358195	30497
14	6	5398928	5398928	46689
15	3	5402692	5402692	73702
16	4	5528454	5528454	56080
17	4	5559021	5559021	52243
18	5	5459029	5459029	33474
19	6	5454606	5454606	23120
20	8	5341977	5341977	31568
21	8	5333679	5333679	33044
22	6	5378068	5378068	23401
23	8	5424819	5424819	42158
24	5	5378604	5378604	20751
25	1	5231305	5231305	5231305
26	3	5415631	5415631	99834
27	4	5398585	5398585	69775
28	3	5503468	5503468	214273
29	6	5449253	5449253	30988
30	4	5489680	5489680	68643
31	5	5258349	5258349	34026
32	3	5241383	5241383	135089
33	6	5304456	5304456	37164
34	4	5264889	5264889	38099
35	4	5387397	5387397	54988
36	4	5455018	5455018	35366
37	5	5437748	5437748	36552
38	3	5124109	5124109	156833
39	2	5299802	5299802	76185
40	3	5249209	5249209	95110

Table 2.4: Polished assembly statistics. Summary statistics for genome assemblies of all isolates after correction with nanopolish

sequencing with assembly approach had the potential to place 20 (91%) of these patients on effective therapy sooner than did standard AST methods.

Overall, the median time to effective antibiotic therapy for the 28 patients infected with carbapenem-resistant *K. pneumoniae* was 61 h (interquartile range [IQR], 43 to 82 h). Of the antibiotics evaluated, ceftazidime-avibactam,

Isolate Number	# of contigs	n50 (bp)	longest contig (bp)	shortest contig (bp)
1	4	5381487	5381487	51167
2	6	5364382	5364382	34310
3	4	5311108	5311108	8844
4	3	4014762	4014762	39294
5	2	5356372	5356372	144603
6	4	5515766	5515766	57137
7	6	5503149	5503149	18504
8	4	5293162	5293162	29283
9	5	5535411	5535411	37975
10	2	5241999	5241999	221833
11	2	5252483	5252483	203962
12	10	3294406	3294406	22763
13	5	5359242	5359242	30542
14	6	5397042	5397042	46264
15	3	5402918	5402918	73707
16	4	5519685	5519685	55829
17	4	5553839	5553839	52295
18	5	5453909	5453909	33605
19	6	5451575	5451575	23120
20	8	5347923	5347923	31560
21	8	5327860	5327860	33040
22	6	5372314	5372314	23510
23	8	5427621	5427621	41250
24	5	5378838	5378838	20779
25	1	5232737	5232737	5232737
26	3	5422958	5422958	100080
27	4	5405798	5405798	69800
28	3	5507869	5507869	214529
29	6	5448433	5448433	31138
30	4	5485433	5485433	68562
31	5	5260434	5260434	34043
32	3	5246142	5246142	134808
33	6	5307245	5307245	37210
34	4	5268031	5268031	37940
35	4	5387449	5387449	54983
36	4	5437773	5437773	35369
37	5	5437459	5437459	36689
38	3	5124074	5124074	156818
39	2	5299231	5299231	76148
40	3	5240201	5240201	94835

Table 2.5: Short-read correction assembly statistics. Summary statistics for genome assemblies of all isolates after correction with short-read data

extended-infusion meropenem, aminoglycosides, fluoroquinolones, polymixins, and tigecycline are generally considered reasonable treatment options for infections caused by carbapenem-resistant *K. pneumoniae* isolates, if active in vitro. As results from the real-time analysis approach were generally available

Isolate Number	# of contigs	n50 (bp)	longest contig (bp)	shortest contig (bp)
1	6	3279607	3279607	39430
2	10	5316522	5316522	14713
3	5	4527157	4527157	16841
4	3	5368538	5368538	17428
5	2	5293767	5293767	129043
6	5	5455776	5455776	14379
7	8	5431055	5431055	17297
8	4	5237837	5237837	27046
9	6	5484734	5484734	34263
10	2	5172417	5172417	209465
11	2	5192109	5192109	198529
12	6	5350919	5350919	26469
13	5	5297830	5297830	33964
14	8	5339196	5339196	21363
15	5	5331989	5331989	9743
16	4	5477999	5477999	36255
17	12	2025764	2123455	2453
18	4	5411221	5411221	51750
19	6	5394690	5394690	16798
20	7	5284029	5284029	23318
21	7	5500581	5500581	29080
22	4	5311389	5311389	24974
23	7	5369624	5369624	30029
25	1	5176691	5176691	5176691
26	3	5322531	5322531	41651
28	5	5442403	5442403	4430
29	8	4425498	4425498	29802
30	5	5440840	5440840	13982
31	5	5192492	5192492	33746
32	4	5162118	5162118	54490
33	4	5236343	5236343	49389
34	2	5210236	5210236	233404
35	3	5325108	5325108	144054
36	4	5380945	5380945	34240
37	5	5373953	5373953	23643
38	3	5068471	5068471	143600
39	2	5239676	5239676	97726
40	3	4418684	4418684	304407

Table 2.6: Rapid assembly statistics. Summary statistics for genome assemblies of all isolates using downsampled data

within 8 h from subcultured isolates, Nanopore real-time analysis could have led to an average time to effective therapy of 41 h (IQR, 33 to 44). Using the Nanopore assembly approach, time to effective therapy could have been reduced to 35 h (IQR, 32 to 42). The time to effective therapy was reduced with the assembly approach compared to that with the real-time approach

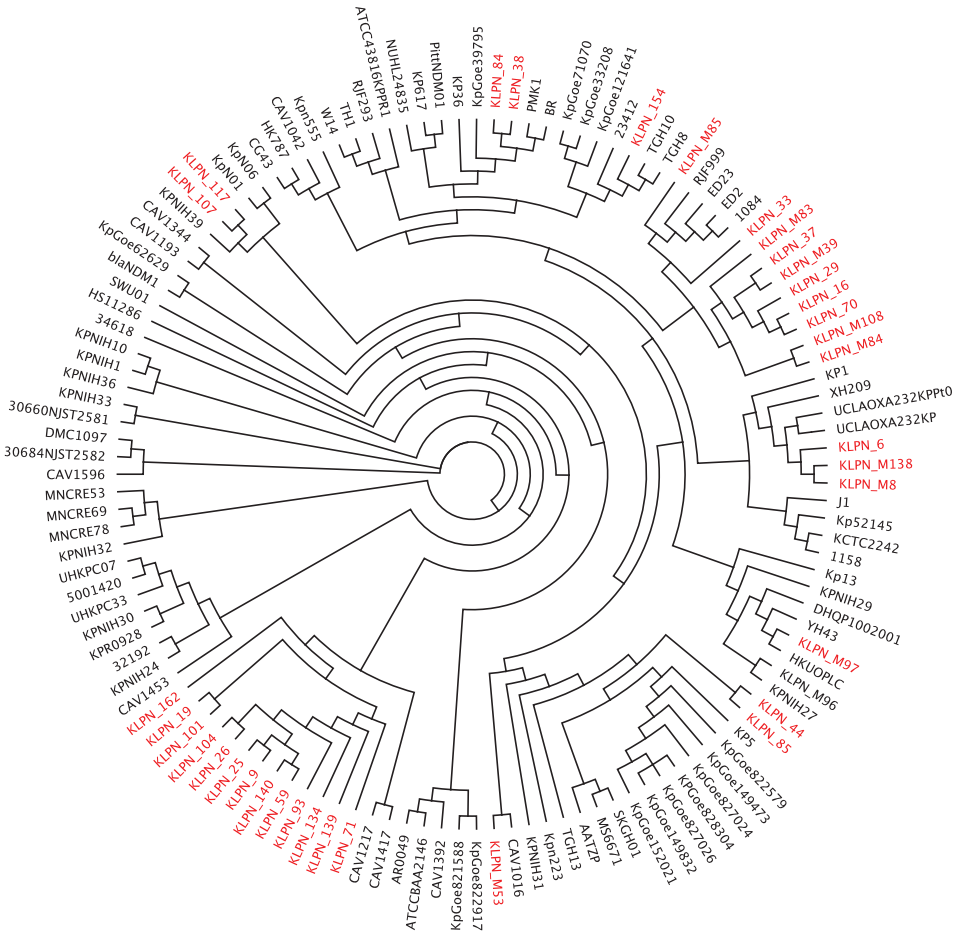

Figure 2.6: Phylogenetic tree of *K. pneumoniae* genomes. Isolates from this study are shown in red, and chromosome-level genomes obtained from GenBank are shown in black.

because the former provided more comprehensive data to infer AST activity (e.g., aminoglycoside resistance, colistin resistance, etc.) than the real-time approach, for which there were delays in antibiotic optimization while awaiting additional AST results. Both the real-time and assembly approaches were significantly faster than the standard approach.

	Phenotypic antimicrobial susceptibility testing results (%)		% agreement with antimicrobial susceptibility testing results		
Antibiotic	Susceptible	Not susceptible	Real-time approach	Assembly approach	Hybrid assembly
Piperacillin-tazobactam	25	75	80	85	85
Ceftriaxone	25	75	93	95	95
Cefepime	28	72	95	98	98
Ceftazidime-avibactam	93	7	100	100	100
Ertapenem	78	22	83	85	85
Meropenem	40	60	93	95	95
Amikacin	78	22	78	85	85
Gentamicin	60	40	45	93	95
Ciprofloxacin	33	67	30	98	98
Colistin	93	7	93	98	98
Doxycycline	50	50	63	80	80
Trimethoprim-sulfamethoxazole	35	65	68	93	93
Overall agreement			77	92	92

Table 2.7: WGS vs phenotypic AST. Percent agreement between three different sequencing and analysis approaches compared to phenotypic antimicrobial susceptibility testing results for 40 *Klebsiella pneumoniae* clinical isolates

2.4 Discussion

Our results demonstrate that Nanopore sequencing can be employed to both accurately and rapidly predict phenotypic AST profiles. Our study builds off previously published proof-of-concept or early insight studies applying Nanopore sequencing for the detection of antimicrobial resistance genes (Judge et al., 2015; Judge et al., 2016; Li et al., 2019; Helm et al., 2017; Xia et al., 2017; Neuert et al., 2018; Hasman et al., 2014; Stoesser et al., 2013). We found an overall agreement of 92% between genotypic results and comprehensive AST results using a Nanopore assembly-based approach. We further demonstrated that assembly-based approaches enhance the ability to identify chromosomal mutations and allelic variants compared to that of

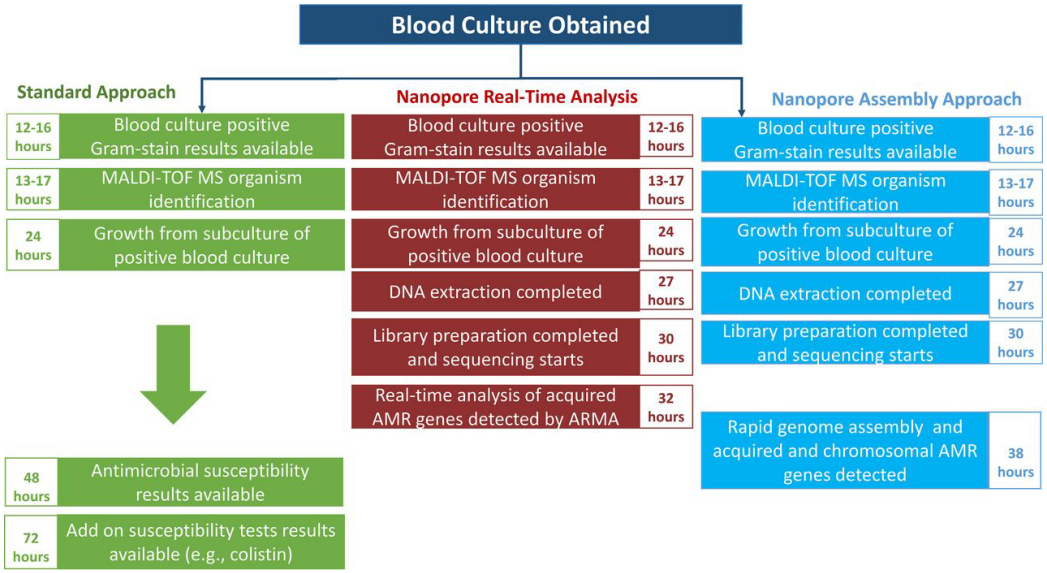

Figure 2.7: Estimated timelines of resistance detection. Schematic of Nanopore sequencing with a real-time analysis and assembly-based approach for identifying resistance genes compared to standard of care testing, using an example of a positive blood culture. MALDI-TOF MS, matrix-assisted laser desorption ionization<U+2013>time of flight mass spectrometry; AMR, antimicrobial resistance; AST, antimicrobial susceptibility testing.

the Nanopore real-time approach. A future method centered on real-time alignment and consensus of raw reads to a database of antimicrobial resistance (AMR) genes will enhance the capabilities of the real-time approach. Nevertheless, even with its current limitations, the Nanopore real-time analysis approach correctly predicted the presumed activity of a number of antibiotics commonly prescribed for Gram-negative infections, such as β-lactams, trimethoprim-sulfamethoxazole, tetracyclines, with reasonable accuracy, with full annotation of acquired AMR genes within 8 h from the time of cultured isolates. Within 14 h of cultured isolates, a Nanopolished assembly-based approach would allow for predictions of the activity of additional agents, such

as fluoroquinolones, aminoglycosides, and polymixins (beyond *mcr*-1 and its variants), due to the ability to detect chromosomal mutations or allelic variants leading to resistance. We believe that with rapid extraction and library preparation techniques, the turn-around times could further be reduced to <3 h for a real-time approach and <9 h for a Nanopore assembly-based approach. Moreover, using a hypothetical trial design, we found that a real-time approach and assembly approach could shorten the average time to effective antibiotic therapy for carbapenem-resistant *K. pneumoniae* infections by 20 h and 26 h, respectively, compared to standard approaches.

Although there have been other investigations of the use of WGS to predict AST results for Gram-negative organisms based on resistance determinants (Shelburne et al., 2017; Cao et al., 2016; Schmidt et al., 2017; Lemon et al., 2017; Judge et al., 2015; Judge et al., 2016; Li et al., 2019; Helm et al., 2017; Xia et al., 2017; Stoesser et al., 2013), ours is the first to use a Nanopore assembly approach for evaluating a broad range of acquired resistance genes and chromosomal mutations in predicting susceptibility results for a comprehensive panel of antibiotics. Previous studies applying Nanopore sequencing for resistance gene detection have applied this methodology to small numbers of isolates and limited evaluations to acquired resistance genes. Furthermore, the potential impact of WGS in reducing time to appropriate antibiotic therapy for highly drug-resistant organisms has not been previously reported. As the science of WGS continues to evolve, the costs of sequencing are becoming more affordable, and the time requirements for DNA extraction, library preparation, assembly, and detection are anticipated to be further reduced. We believe

that this methodology can accurately expedite antibiotic decision making for critically ill patients infected with MDRGN organisms. WGS also provides the ability to identify emerging mechanisms of resistance, such as *mcr* variants or novel mechanisms of resistance against newly approved agents (Haidar et al., 2017b; Haidar et al., 2017a).

As we continue to gain further insights into the complexities of resistance mechanisms present in Gram-negative organisms, it is becoming increasingly clear that sole reliance on the detection of β-lactamase genes to identify antibiotic resistance is insufficient. Such is the case with commercially available PCR-based methodologies (e.g., Cepheid Xpert Carba-R assay, BioFire FilmArray blood culture identification panel, Verigene Gram-negative blood culture test, etc.) that fail to identify the wide array of additional resistance determinants (e.g., porin deletions, multidrug efflux pumps, functioning of the two-component regulatory system, DNA gyrase mutations, off-panel targets, etc.). Using carbapenem-resistant *K. pneumoniae* as a case study, over half of isolates render carbapenem antibiotics ineffective due to non-carbapenemase-mediated mechanisms (Tamma et al., 2017). Similarly, carbapenem resistance among *Pseudomonas aeruginosa* strains in the United States is predominantly mediated by noncarbapenemase mechanisms, including the loss of OprD porin expression and/or upregulation of MexAB-OprM efflux pumps (Lister, Wolter, and Hanson, 2009). WGS offers a more comprehensive approach to identifying clinically relevant antibiotic resistance compared with standard PCR technologies.

Nanopore technology offers real-time DNA sequencing of long reads,

which facilitate the process of high-quality genome assembly (Lu, Giordano, and Ning, 2016). Additionally, they are better able to capture large regions of structural variation (e.g., insertions, deletions, duplications, translocations, inversions, etc.) and resolve repetitive regions accurately, compared to technologies which generate short-read data. However, there are some drawbacks to Nanopore sequencing (Lu, Giordano, and Ning, 2016). Chromosomal mutations as small as single-base polymorphisms in DNA can result in drastic changes to proteins due to frame shifts or early truncations. Indels and resulting false positives in chromosomal mutations can be difficult to distinguish with Nanopore sequencing alone (Jain et al., 2018). The assembly process generally removes random error, given sufficient sequencing coverage. However, systematic errors in the form of homopolymer indels and methylation errors still result in disagreement compared with Illumina-corrected assemblies. Nanopolish corrects most homopolymer indels, improving accuracy, but in its current form, it is not all-inclusive (Loman, Quick, and Simpson, 2015). Methylation motifs alter the electrical signal from Nanopore sequencing, resulting in errors in base calling. We anticipate further improvements to the quality of Nanopore sequencing and available bioinformatics, enabling this technology to rapidly and accurately resolve variants of acquired resistance genes and characterize chromosomal mutations in the absence of assembly.

We assessed the minimum sequencing time to be able to accurately predict AST using Nanopore sequencing. We determined that a minimum of 10 reads per gene were required and that a sequencing run of 1 h is sufficient to predict the AST profile, using either a real-time sequencing workflow or an assembly

approach. In fact, we were able to detect 10 reads of all AMR genes within 14 min of starting the sequencing run and 40 reads of each gene within 1 h. This is an important quality control measure as we begin to consider these methods for clinical application.

There are several remaining limitations to this work. First, there were some antimicrobial susceptibility profiles for which we were unable to identify the associated resistance mechanisms, suggesting that our approach to resistance detection was not comprehensive. As an example, we were only able to accurately predict piperacillin-tazobactam activity 85% of the time. This is similar to what others have shown (Shelburne et al., 2017), and it is likely due to efflux pumps, where expression analysis is required. Second, there were some agents tested for which no resistance was observed (e.g., tigecycline), precluding us from including these agents in our analysis. Third, we only evaluated *K. pneumoniae* isolates from a single region of the United States. Validation needs to occur in larger data sets with a more diverse array of genera and species and resistance mechanisms, including some not observed in our isolates, such as the *mcr* genes. Additionally, as antibiotic susceptibility criteria include several considerations, such as wild-type MIC distributions, pharmacokinetic-pharmacodynamic modeling, clinical outcomes data, etc., it can be challenging to determine the most accurate MIC that signifies nonsusceptibility. The European Committee on Antimicrobial Susceptibility Testing suggests that epidemiologic cutoff values (i.e., wild-type versus non-wild-type distributions) are the preferred values for correlation with resistance genes (Eucast, n.d.). We elected to use antibiotic breakpoints, as these remain the

most relevant metric for quantifying antibiotic resistance for clinicians, but we realize that this might not be the most accurate proxy.

In conclusion, we were able to leverage the long reads, rapid turnaround time, and real-time analytic capabilities of Nanopore sequencing to accurately identify resistance loci. With the continued rise in highly drug-resistant infections, the need for rapid and accurate methods to detect antibiotic resistance is becoming increasingly important. Continued enhancements to WGS may permit real-time AMR gene detection from clinical isolates in the near future.

2.5 Materials and Methods

2.5.1 Study cohort

Forty clinical *Klebsiella pneumoniae* complex isolates collected between 2016 and 2017 and processed at the Johns Hopkins Hospital Medical Microbiology Laboratory were included in the present study. Isolates for inclusion were deliberately selected based on their diversity of AST results and mechanisms of resistance and included 30 carbapenem-resistant (21 carbapenemase producers and 10 non-carbapenemase producers) and 9 carbapenem-susceptible *K. pneumoniae* isolates. As this was a proof-of-concept study comparing different WGS approaches, we decided to focus on a single genus and species (i.e., *K. pneumoniae*) to increase the number of isolates with any single resistance mechanism identified, as some resistance mechanisms are genus and species specific, especially among chromosomal mutations leading to resistance. Isolates were subcultured from frozen stock to tryptic soy agar (TSA) with 5% blood agar. A second subculture was performed prior to DNA extraction.

K. pneumoniae isolates from deep tissue (n = 3), intraabdominal fluid (n = 1), blood (n = 10), respiratory (n = 10), and urine (n = 16) were included.

2.5.2 Species and antimicrobial susceptibility testing

Bacterial genus and species were identified using matrix-assisted laser desorption ionization–time of flight mass spectrometry (Bruker Daltonics Inc., Billerica, MA). AST results were determined using the BD Phoenix Automated System NMIC-303 panels (BD Diagnostics, Sparks, Maryland). MICs were confirmed by broth microdilution (BMD) with Sensititre GNX2F Gram-negative panels (Thermo Fisher Scientific, Indianapolis, IN) and the ceftazidime-avibactam Etest (bioMérieux, France) (Matuschek et al., 2018). BMD was repeated on isolates with 2-fold or greater MIC discrepancies between the Phoenix automated panel and initial BMD results. Interpretive criteria established by the Clinical and Laboratory Standards Institute (CLSI) were used to define antibiotic susceptibility. AST results in both the "intermediate" and "resistant" ranges were categorized as resistant.

2.5.3 Whole-genome sequencing and antimicrobial resistance gene detection

As Nanopore sequencing has a substantial raw sequencing error rate, three separate sequencing and analysis pipelines were performed, as follows: (i) a real-time Nanopore (Oxford, England) analysis approach that identified acquired

and chromosomal resistance genes by applying Metrichor's Antimicrobial Resistance Mapping Application (ARMA) (https://nanoporetech.com/resource-centre/real-time-detection-antibiotic-resistance-genes-using-oxford-nanopore-technologies); (ii) an assembly-based Nanopore approach that uses Canu and Nanopolish (Loman, Quick, and Simpson, 2015; Koren et al., 2017) to compute high-identity sequences with identification of resistance genes, using Abricate ResFinder results as well as chromosomal mutations (e.g., ompK35 and ompK36 mutations, *gyrA* and *parC* mutations, etc.) from genomic assemblies, using minimap2 (Li, 2018) and a tool evaluating the impact on amino acid translation based on resulting codon changes (https://github.com/gpertea/pwasm); and (iii) a Pilon-corrected hybrid approach using both Illumina (Illumina, San Diego, California) and Nanopore sequencing (Walker et al., 2014). The short-read correction of Nanopore assemblies served as the reference standard to determine the accuracy of Nanopore sequencing results.

Long-read genomic sequencing was performed using the third generation Oxford Nanopore MinION MkIb (Isolates 1-30) and GridION X5 (Isolates 31-40; Oxford, England) sequencing instruments. Genomic DNA was extracted from pure cultures using the DNeasy PowerBiofilm Kit (Qiagen, Hilden, Germany). Each Nanopore sequencing library was prepared using 5 µg of DNA with the 1D ligation kit (SQK-LSK108, Oxford Nanopore Technologies) using R9.4 flowcells (FLO-MIN106). A single isolate was run per flowcell. MinKNOW software was used to collect sequencing data.

For the real-time analysis approach, the Oxford Nanopore Technology "What's In My Pot" (WIMP) (Juul et al., 2015) and Metrichor's Antimicrobial

Resistance Mapping Application (ARMA) real-time analysis tools were run in parallel to both sequencing and base-calling. WIMP uses a customized pipeline applying the Centrifuge metagenomic classifier to the sequencing reads. ARMA aligns sequencing reads to the Comprehensive Antimicrobial Database (CARD) (Jia et al., 2017), identifying antimicrobial resistance (AMR) gene hits.

For the assembly-based approach, Albacore v2.1.3 was used to base-call. Raw data were corrected, trimmed, and assembled using Canu v1.6 (Koren et al., 2017) using default parameters. Genome size was assumed to be 5.3Mb. For isolates that did not assemble under default parameters, either read quality restrictions were relaxed or minimum overlap requirements were shortened as suggested by the software documentation. Assembled contigs were screened for resistance genes using Abricate (https://github.com/tseemann/abricate), a tool that uses alignment to search for resistance gene sequences from several database including ResFinder, CARD, ARG-ANNOT, and the NCBI AMR Reference Gene Database. ResFinder results were evaluated in this study (Zankari et al., 2012). In a separate experiment, we were able to build high quality genomes from nanopore electrical data in under 6 hours using a machine with 36 cores and 72GB RAM. Thirteen random blocks of 4000 reads were sampled and taken through basecalling, assembly and polishing. Conservatively, it takes at most an hour to reach 52,000 reads on a typical run.

WGS was also conducted using Illumina MiSeq short-read sequencing to increase assembly accuracy (Illumina, San Diego, California). A drawback with Illumina sequencing is the turn-around time as sequencing alone requires

between 19-24 hours for 300 cycles. As the ultimate goal would be to use Nanopore sequencing alone for both resistance gene and chromosomal mutation identification, in the current analysis, the Pilon-corrected assemblies were not meant to supplant or supplement Nanopore sequencing but rather to serve as a reference standard to determine the accuracy of Nanopore sequencing results. Approximately 100-500 ng of gDNA was used to prepare sequencing libraries using the Nextera Flex Kit. The each Illumina library was then sequenced using both v2 2x150 and v3 2x75 reagents on an Illumina Miseq. These reads were used to correct the more error prone Nanopore assembly with the Pilon v1.22 software package in conjunction with the short-read aligner Bowtie2 v2.2.6. Phylogenetic trees were generated using Parsnp v1.2 (Treangen et al., 2014).

2.5.4 Predicted correlations between WGS and AST results

Predictions of resistance were performed without reference to phenotypic data, unless otherwise noted in Results. The correlations of resistance genes and sequence variants with anticipated AST results for the evaluated antibiotics were determined based on reference gene databases and the published literature (Bush and Jacoby, 2010; Bush, Jacoby, and Medeiros, 1995; Jacoby, 2009; Naas, Poirel, and Nordmann, 2008; Evans and Amyes, 2014; Zhang et al., 2014; Feng-Jui et al., 2003; Vetting et al., 2011; Kim et al., 2009; Ramirez and Tolmasky, 2010; Galimand, Courvalin, and Lambert, 2012; Liakopoulos, Mevius, and Ceccarelli, 2016; Tärnberg, Nilsson, and Monstein, 2009; Bonnet, 2004; Roberts, 2005; Fevre et al., 2005; Fu et al., 2013; Poirel, Naas, and Nordmann,

2010). Antimicrobial resistance genes and mutation results were reviewed manually to predict AST results. For associations for which ambiguity existed, the following rules were established a priori:

- SHV 2be extended-spectrum β-lactamase enzymes were assumed to inactivate ceftriaxone and cefepime if they contained G238S or E240K mutations (Liakopoulos, Mevius, and Ceccarelli, 2016; Howard Christopher et al., 2002)

- The relative contributions of porins OmpK35 and/or OmpK36 truncations to β-lactam resistance have not been well established (Zhang et al., 2014; Shelburne et al., 2017; Tsai et al., 2011; Landman, Bratu, and Quale, 2009; Jiang et al., 2009; Doménech-Sánchez et al., 2003). There appears to be consensus that when premature stop codons are present for both, in conjunction with ESBLs or AmpC cephalosporinases, carbapenem resistance is likely.

- Aminoglycoside modifying enzymes (AMEs)- including aminoglycoside N82 acetyltransferases [AACs], aminoglycoside O-nucleotidyltransferases [ANTs], or aminoglycoside O-phosphotransferases [APHs] are anticipated to confer resistance to gentamicin and tobramycin (Ramirez and Tolmasky, 2010). Correlations between the number of different enzymes produced and the association with aminoglycoside resistance are not well defined. After exploratory analysis, we predicted that when four or more AMEs were present, gentamicin and tobramycin resistance was likely.

- The AME *aac(6')Ib-cr* was also evaluated for its contribution to ciprofloxacin resistance (Ramirez and Tolmasky, 2010).

- Ribosomal RNA methyltransferases (*armA*, *rmtA*, *rmtB*, *rmtC*, *rmtD*, or *rmtE*) are expected to confer resistance to amikacin, gentamicin, and tobramycin (Galimand, Courvalin, and Lambert, 2012).

- Plasmid-encoded quinolone resistance genes *qnrB*, *qnrS*, *aac(6')Ib-cr* and *oqxAB* (chromosomally encoded) were predicted to cause low-level fluoroquinolone resistance, but not frank resistance (Kim et al., 2009; Ruiz, Pons, and Gomes, 2012; Jacoby, 2005).

Single *gyrA* or *parC* mutations were not predicted to cause fluoroquinolone resistance but two-step *gyrA* mutations or the presence of both *gyrA* and *parC* mutations translating to amino acid changes were expected to result in fluoroquinolone resistance (Ruiz, Pons, and Gomes, 2012; Jacoby, 2005).

2.5.5 Clinical data

An infectious diseases physician (P.D.T.) confirmed all isolates to be representative of clinical infections by manual chart review. Detailed clinical and treatment data were collected to identify if the time to effective antibiotic therapy would have decreased had Nanopore sequencing data with a real-time analysis approach or an assembly-based approach been used in predicting AST results compared to that with standard AST identification methods. Dates and times antibiotic regimen changes occurred and the timing of the availability of clinical culture results were recorded. Furthermore, time points for

the steps involved in Nanopore sequencing and identification of resistance determinants were also recorded. The Johns Hopkins University School of Medicine Institutional Review Board approved this study, with a waiver of informed consent.

2.5.6 Data Availability

Sequencing data for this study were deposited in the Sequence Read Archive (BioProject number PRJNA496461).

References

Didelot, Xavier, Rory Bowden, Daniel J Wilson, Tim E A Peto, and Derrick W Crook (2012). "Transforming clinical microbiology with bacterial genome sequencing". en. In: *Nat. Rev. Genet.* 13.9, pp. 601–612.

Shelburne, Samuel A, Jiwoong Kim, Jose M Munita, Pranoti Sahasrabhojane, Ryan K Shields, Ellen G Press, Xiqi Li, Cesar A Arias, Brandi Cantarel, Ying Jiang, Min S Kim, Samuel L Aitken, and David E Greenberg (2017). "Whole-Genome Sequencing Accurately Identifies Resistance to Extended-Spectrum β-Lactams for Major Gram-Negative Bacterial Pathogens". en. In: *Clin. Infect. Dis.* 65.5, pp. 738–745.

Caliendo, Angela M, David N Gilbert, Christine C Ginocchio, Kimberly E Hanson, Larissa May, Thomas C Quinn, Fred C Tenover, David Alland, Anne J Blaschke, Robert A Bonomo, Karen C Carroll, Mary Jane Ferraro, Lisa R Hirschhorn, W Patrick Joseph, Tobi Karchmer, Ann T MacIntyre, L Barth Reller, Audrey F Jackson, and Infectious Diseases Society of America (IDSA) (2013). "Better tests, better care: improved diagnostics for infectious diseases". en. In: *Clin. Infect. Dis.* 57 Suppl 3, S139–70.

Tamma, Pranita D, Katherine E Goodman, Anthony D Harris, Tsigereda Tekle, Ava Roberts, Abimbola Taiwo, and Patricia J Simner (2017). "Comparing the Outcomes of Patients With Carbapenemase-Producing and Non-Carbapenemase-Producing Carbapenem-Resistant *Enterobacteriaceae* Bacteremia". en. In: *Clin. Infect. Dis.* 64.3, pp. 257–264.

Cao, Minh Duc, Devika Ganesamoorthy, Alysha G Elliott, Huihui Zhang, Matthew A Cooper, and Lachlan J M Coin (2016). "Streaming algorithms for identification pathogens and antibiotic resistance potential from real-time MinION™ sequencing". en. In: *Gigascience* 5.1, s13742–016–0137–2.

Schmidt, K, S Mwaigwisya, L C Crossman, M Doumith, D Munroe, C Pires, A M Khan, N Woodford, N J Saunders, J Wain, J O'Grady, and D M Livermore (2017). "Identification of bacterial pathogens and antimicrobial resistance

directly from clinical urines by nanopore-based metagenomic sequencing". en. In: *J. Antimicrob. Chemother.* 72.1, pp. 104–114.

Lemon, Jamie K, Pavel P Khil, Karen M Frank, and John P Dekker (2017). "Rapid Nanopore Sequencing of Plasmids and Resistance Gene Detection in Clinical Isolates". en. In: *J. Clin. Microbiol.* 55.12, pp. 3530–3543.

Judge, Kim, Simon R Harris, Sandra Reuter, Julian Parkhill, and Sharon J Peacock (2015). "Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes". en. In: *J. Antimicrob. Chemother.* 70.10, pp. 2775–2778.

Judge, Kim, Martin Hunt, Sandra Reuter, Alan Tracey, Michael A Quail, Julian Parkhill, and Sharon J Peacock (2016). "Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology". en. In: *Microb Genom* 2.9, e000085.

Li, Ruichao, Miaomiao Xie, Ning Dong, Dachuan Lin, Xuemei Yang, Marcus Ho Yin Wong, Edward Wai-Chi Chan, and Sheng Chen (2019). "Erratum to: Efficient generation of complete sequences of MDR-encoding plasmids by rapid assembly of MinION barcoding sequencing data". en. In: *Gigascience* 8.3.

Helm, Eric van der, Lejla Imamovic, Mostafa M Hashim Ellabaan, Willem van Schaik, Anna Koza, and Morten O A Sommer (2017). "Rapid resistome mapping using nanopore sequencing". en. In: *Nucleic Acids Res.* 45.8, e61.

Xia, Yu, An-Dong Li, Yu Deng, Xiao-Tao Jiang, Li-Guan Li, and Tong Zhang (2017). "MinION Nanopore Sequencing Enables Correlation between Resistome Phenotype and Genotype of Coliform Bacteria in Municipal Sewage". en. In: *Front. Microbiol.* 8, p. 2105.

Neuert, Saskia, Satheesh Nair, Martin R Day, Michel Doumith, Philip M Ashton, Kate C Mellor, Claire Jenkins, Katie L Hopkins, Neil Woodford, Elizabeth de Pinna, Gauri Godbole, and Timothy J Dallman (2018). "Prediction of Phenotypic Antimicrobial Resistance Profiles From Whole Genome Sequences of Non-typhoidal *Salmonella enterica*". en. In: *Front. Microbiol.* 9, p. 592.

Hasman, Henrik, Dhany Saputra, Thomas Sicheritz-Ponten, Ole Lund, Christina Aaby Svendsen, Niels Frimodt-Møller, and Frank M Aarestrup (2014). "Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples". en. In: *J. Clin. Microbiol.* 52.1, pp. 139–146.

Stoesser, N, E M Batty, D W Eyre, M Morgan, D H Wyllie, C Del Ojo Elias, J R Johnson, A S Walker, T E A Peto, and D W Crook (2013). "Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data". en. In: *J. Antimicrob. Chemother.* 68.10, pp. 2234–2244.

Haidar, Ghady, Nathan J Philips, Ryan K Shields, Daniel Snyder, Shaoji Cheng, Brian A Potoski, Yohei Doi, Binghua Hao, Ellen G Press, Vaughn S Cooper, Cornelius J Clancy, and M Hong Nguyen (2017b). "Ceftolozane-Tazobactam for the Treatment of Multidrug-Resistant *Pseudomonas aeruginosa* Infections: Clinical Effectiveness and Evolution of Resistance". en. In: *Clin. Infect. Dis.* 65.1, pp. 110–120.

Haidar, Ghady, Cornelius J Clancy, Ryan K Shields, Binghua Hao, Shaoji Cheng, and M Hong Nguyen (2017a). "Mutations in blaKPC-3 That Confer Ceftazidime-Avibactam Resistance Encode Novel KPC-3 Variants That Function as Extended-Spectrum β-Lactamases". en. In: *Antimicrob. Agents Chemother.* 61.5.

Lister, Philip D, Daniel J Wolter, and Nancy D Hanson (2009). "Antibacterial-resistant *Pseudomonas aeruginosa*: clinical impact and complex regulation of chromosomally encoded resistance mechanisms". en. In: *Clin. Microbiol. Rev.* 22.4, pp. 582–610.

Lu, Hengyun, Francesca Giordano, and Zemin Ning (2016). "Oxford Nanopore MinION Sequencing and Genome Assembly". en. In: *Genomics Proteomics Bioinformatics* 14.5, pp. 265–279.

Jain, Miten, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads". en. In: *Nat. Biotechnol.* 36.4, pp. 338–345.

Loman, Nicholas J, Joshua Quick, and Jared T Simpson (2015). "A complete bacterial genome assembled de novo using only nanopore sequencing data". en. In: *Nat. Methods* 12.8, pp. 733–735.

Eucast, Wgs (n.d.). *Subcommitte Consultation on Report from the EUCAST Subcommittee on the Role of Whole Genome Sequencing (WGS) in Antimicrobial Susceptibility Testing of Bacteria*.

Matuschek, Erika, Jenny Åhman, Catherine Webster, and Gunnar Kahlmeter (2018). "Antimicrobial susceptibility testing of colistin–evaluation of seven commercial MIC products against standard broth microdilution for *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, and *Acinetobacter* spp." In: *Clin. Microbiol. Case Rep.*

Koren, Sergey, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy (2017). "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation". en. In: *Genome Res.* 27.5, pp. 722–736.

Li, Heng (2018). "Minimap2: pairwise alignment for nucleotide sequences". en. In: *Bioinformatics* 34.18, pp. 3094–3100.

Walker, Bruce J, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, and Ashlee M Earl (2014). "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement". en. In: *PLoS One* 9.11, e112963.

Juul, Sissel, Fernando Izquierdo, Adam Hurst, Xiaoguang Dai, Amber Wright, Eugene Kulesha, Roger Pettett, and Daniel J Turner (2015). "What's in my pot? Real-time species identification on the MinION™". en.

Jia, Baofeng, Amogelang R Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K Tsang, Briony A Lago, Biren M Dave, Sheldon Pereira, Arjun N Sharma, Sachin Doshi, Mélanie Courtot, Raymond Lo, Laura E Williams, Jonathan G Frye, Tariq Elsayegh, Daim Sardar, Erin L Westman, Andrew C Pawlowski, Timothy A Johnson, Fiona S L Brinkman, Gerard D Wright, and Andrew G McArthur (2017). "CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database". en. In: *Nucleic Acids Res.* 45.D1, pp. D566–D573.

Zankari, Ea, Henrik Hasman, Salvatore Cosentino, Martin Vestergaard, Simon Rasmussen, Ole Lund, Frank M Aarestrup, and Mette Voldby Larsen (2012). "Identification of acquired antimicrobial resistance genes". en. In: *J. Antimicrob. Chemother.* 67.11, pp. 2640–2644.

Treangen, Todd J, Brian D Ondov, Sergey Koren, and Adam M Phillippy (2014). "The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes". In: *Genome Biol.* 15.11, p. 524.

Bush, Karen and George A Jacoby (2010). "Updated functional classification of beta-lactamases". en. In: *Antimicrob. Agents Chemother.* 54.3, pp. 969–976.

Bush, K, G A Jacoby, and A A Medeiros (1995). "A functional classification scheme for beta-lactamases and its correlation with molecular structure". en. In: *Antimicrob. Agents Chemother.* 39.6, pp. 1211–1233.

Jacoby, George A (2009). "AmpC beta-lactamases". en. In: *Clin. Microbiol. Rev.* 22.1, 161–82, Table of Contents.

Naas, T, L Poirel, and P Nordmann (2008). "Minor extended-spectrum beta-lactamases". en. In: *Clin. Microbiol. Infect.* 14 Suppl 1, pp. 42–52.

Evans, Benjamin A and Sebastian G B Amyes (2014). "OXA β-lactamases". en. In: *Clin. Microbiol. Rev.* 27.2, pp. 241–263.

Zhang, Ying, Xiaofei Jiang, Yanyan Wang, Gang Li, Yueru Tian, Hong Liu, Fuqi Ai, Yiming Ma, Bei Wang, Feiyi Ruan, and Kumar Rajakumar (2014). "Contribution of β-lactamases and porin proteins OmpK35 and OmpK36 to carbapenem resistance in clinical isolates of KPC-2-producing *Klebsiella pneumoniae*". en. In: *Antimicrob. Agents Chemother.* 58.2, pp. 1214–1217.

Feng-Jui, Chen, Tsai-Ling Lauderdale, Monto Ho, and Hsiu-Jung Lo (2003). "The Roles of Mutations in gyrA, parC, and ompK35 in Fluoroquinolone Resistance in *Klebsiella pneumoniae*". en. In: *Microbial Drug Resistance; New Rochelle* 9.3, pp. 265–271.

Vetting, Matthew W, Subray S Hegde, Minghua Wang, George A Jacoby, David C Hooper, and John S Blanchard (2011). "Structure of QnrB1, a plasmid-mediated fluoroquinolone resistance factor". en. In: *J. Biol. Chem.* 286.28, pp. 25265–25273.

Kim, Hong Bin, Minghua Wang, Chi Hye Park, Eui-Chong Kim, George A Jacoby, and David C Hooper (2009). "oqxAB encoding a multidrug efflux pump in human clinical isolates of *Enterobacteriaceae*". en. In: *Antimicrob. Agents Chemother.* 53.8, pp. 3582–3584.

Ramirez, Maria S and Marcelo E Tolmasky (2010). "Aminoglycoside modifying enzymes". en. In: *Drug Resist. Updat.* 13.6, pp. 151–171.

Galimand, Marc, Patrice Courvalin, and Thierry Lambert (2012). "RmtF, a new member of the aminoglycoside resistance 16S rRNA N7 G1405 methyltransferase family". en. In: *Antimicrob. Agents Chemother.* 56.7, pp. 3960–3962.

Liakopoulos, Apostolos, Dik Mevius, and Daniela Ceccarelli (2016). "A Review of SHV Extended-Spectrum β-Lactamases: Neglected Yet Ubiquitous". en. In: *Front. Microbiol.* 7, p. 1374.

Tärnberg, Nilsson, and Monstein (2009). "Molecular identification of blaSHV, blaLEN and blaOKP β-lactamase genes in *Klebsiella pneumoniae* by bi-directional sequencing of universal SP6-and T7 ..." In: *Mol. Cell. Probes*.

Bonnet, R (2004). "Growing group of extended-spectrum beta-lactamases: the CTX-M enzymes". en. In: *Antimicrob. Agents Chemother.* 48.1, pp. 1–14.

Roberts, Marilyn C (2005). "Update on acquired tetracycline resistance genes". en. In: *FEMS Microbiol. Lett.* 245.2, pp. 195–203.

Fevre, Cindy, Virginie Passet, François-Xavier Weill, Patrick A D Grimont, and Sylvain Brisse (2005). "Variants of the *Klebsiella pneumoniae* OKP chromosomal beta-lactamase are divided into two main groups, OKP-A and OKP-B". en. In: *Antimicrob. Agents Chemother.* 49.12, pp. 5149–5152.

Fu, Yingmei, Wenli Zhang, Hong Wang, Song Zhao, Yang Chen, Fanfei Meng, Ying Zhang, Hui Xu, Xiaobei Chen, and Fengmin Zhang (2013). "Specific patterns of gyr A mutations determine the resistance difference to ciprofloxacin and levofloxacin in *Klebsiella pneumoniae* and *Escherichia coli*". en. In: *BMC Infect. Dis.* 13.1, pp. 1–6.

Poirel, Laurent, Thierry Naas, and Patrice Nordmann (2010). "Diversity, epidemiology, and genetics of class D beta-lactamases". en. In: *Antimicrob. Agents Chemother.* 54.1, pp. 24–38.

Howard Christopher, van Daal Angela, Kelly Gregory, Schooneveldt Jacqueline, Nimmo Graeme, and Giffard Philip M. (2002). "Identification and Minisequencing-Based Discrimination of SHV β-Lactamases in Nosocomial Infection-Associated *Klebsiella pneumoniae* in Brisbane, Australia". In: *Antimicrob. Agents Chemother.* 46.3, pp. 659–664.

Tsai, Yu-Kuo, Chang-Phone Fung, Jung-Chung Lin, Jiun-Han Chen, Feng-Yee Chang, Te-Li Chen, and L Kristopher Siu (2011). "*Klebsiella pneumoniae* outer membrane porins OmpK35 and OmpK36 play roles in both antimicrobial resistance and virulence". en. In: *Antimicrob. Agents Chemother.* 55.4, pp. 1485–1493.

Landman, David, Simona Bratu, and John Quale (2009). "Contribution of OmpK36 to carbapenem susceptibility in KPC-producing *Klebsiella pneumoniae*". en. In: *J. Med. Microbiol.* 58.Pt 10, pp. 1303–1308.

Jiang, Xiuhong, Bjorn A Espedido, Sally R Partridge, Lee C Thomas, Feng Wang, and Jonathan R Iredell (2009). "Paradoxical effect of *Klebsiella pneumoniae* OmpK36 porin deficiency". en. In: *Pathology* 41.4, pp. 388–392.

Doménech-Sánchez, Antonio, Luis Martínez-Martínez, Santiago Hernández-Allés, María del Carmen Conejo, Alvaro Pascual, Juan M Tomás, Sebastián Albertí, and Vicente Javier Benedí (2003). "Role of *Klebsiella pneumoniae* OmpK35 porin in antimicrobial resistance". en. In: *Antimicrob. Agents Chemother.* 47.10, pp. 3332–3335.

Ruiz, Joaquim, Maria J Pons, and Cláudia Gomes (2012). "Transferable mechanisms of quinolone resistance". en. In: *Int. J. Antimicrob. Agents* 40.3, pp. 196–203.

Jacoby, George A (2005). "Mechanisms of resistance to quinolones". en. In: *Clin. Infect. Dis.* 41 Suppl 2, S120–6.

Chapter 3

Genome assembly of *Candida nivariensis*

Portions of this chapter originally appeared in:

Fan Y, Gale AN, Bailey A, Barnes K, Colotti K, Mass M, et al. Genome and transcriptome of a pathogenic yeast, *Candida nivariensis*. G3 Genes | Genomes | Genetics. 2021;11. doi:10.1093/g3journal/jkab137

3.1 Abstract

We present a highly contiguous genome and transcriptome of the pathogenic yeast, *Candida nivariensis*. We sequenced both the DNA and RNA of this species using both the Oxford Nanopore Technologies and Illumina platforms. We assembled the genome into an 11.8Mb draft composed of 16 contigs with an N50 of 886 Kb, including a circular mitochondrial sequence of 28 Kb. Using direct RNA nanopore sequencing and Illumina cDNA sequencing, we constructed an annotation of our new assembly, supplemented by lifting over genes from *Saccharomyces cerevisiae* and *Candida glabrata*.

3.2 Introduction

For immunocompromised hosts, opportunistic infections caused by drug-resistant fungi of the *Candida* genus are a major source of morbidity and mortality (Borman et al., 2008). In particular, *Candida nivariensis*, a close relative to *Candida glabrata*, has emerged in recent years as especially resistant to antifungal therapies (Borman et al., 2008). However, due to its phenotypic similarities to *C. glabrata*, *C. nivariensis* is generally underidentified and easily misdiagnosed, and currently, only molecular approaches can distinguish the two (Aznar-Marin et al., 2016), spurring whole-genome sequencing studies on the clade (Gabaldón et al., 2013).

Accurate assembly of repetitive genomic regions is crucial for understanding genetic diversity and virulence in pathogenic species. Fungal pathogens have long been known to exhibit a high degree of genome plasticity to enhance fitness in various environments (Croll, Zala, and McDonald, 2013; Ford et al., 2015; López-Fuentes et al., 2018; Carreté et al., 2019; Todd et al., 2019). Repetitive subtelomeric regions in particular play a crucial role in virulence for many pathogenic organisms (Barry et al., 2003; De Las Peñas et al., 2003). Many yeasts' subtelomeric regions contain and regulate the expression of genes crucial for biofilm formation, carbohydrate utilization, and cellular adhesion (Naumov, Naumova, and Louis, 1995; De Las Peñas et al., 2003; Iraqui et al., 2005). These gene families often undergo rapid evolution through changes in copy number and sequence through either SNPs or indels (Carreto et al., 2008; Brown, Murray, and Verstrepen, 2010; Anderson et al., 2015). However, these subtelomeric regions remain one of the most difficult sections of the genome to

accurately assemble due to their repetitive nature and high sequence similarity between genes, making genetic analysis cumbersome (Brown, Murray, and Verstrepen, 2010).

One of the gene families of great interest to the pathogenic yeast field are the GPI-anchored cell wall proteins. This protein family includes many genes that encode for adhesion proteins that are found in various members of the *Candida* genus, and play a key role in pathogenicity, being involved in regulation of biofilm formation, cell-to-cell contact, and host-pathogen interactions (Timmermans et al., 2018; McCall et al., 2019). With the many roles these genes play in infection, the accurate identification and understanding of the genetic variation of these genes is vital to combating fungal pathogens.

Unfortunately, like many eukaryotic pathogens, the current reference genome for *C. nivariensis* (GenBank: GCA_001046915.1) is highly fragmented. Constructed from sequencing of strain CBS9983, the reference genome consists of 123 contigs with an N50 of 248Kb (Gabaldón et al., 2013), meaning that at least half of the total genome length is contained in contigs 248Kb or longer. This is typical of genomes assembled from limited short-read sequencing data; though short reads are highly accurate, assembling them into contiguous genomes is challenging depending on the size and complexity of the genome. Such short read assemblies have limited utility since large scale variants, repetitive regions, and genome structure remain difficult to elucidate, though they are often involved in the genome plasticity of pathogenic yeasts (Carreté et al., 2018). In contrast, long-read sequencing data has been shown to produce much more contiguous assemblies, and have been crucial

in sequencing through large repetitive regions, as well as assessing structural variants. However, read accuracy on the ONT platform in particular ranges from 86% for early basecaller versions (Wick, Judd, and Holt, 2019) to 97% as currently reported by ONT. This is lower than the read accuracy of short-read Illumina sequencing, which achieves 99.9% accuracy (Fox et al., 2014). In consensus sequences, most random errors can be corrected by other reads covering the same genomic loci, resulting in >99% consensus accuracy (Wick, Judd, and Holt, 2019). However, systematic errors occurring in most or all of the reads cannot be corrected this way. For ONT data, indels at homopolymers dominate systematic errors (Wick, Judd, and Holt, 2019). These persistent errors can be problematic for gene prediction and annotation in downstream analysis (Watson and Warr, 2019) and are typically corrected with more accurate short-read data in mappable regions (Garrison and Marth, 2012; Walker et al., 2014; Vaser et al., 2017).

Having a genome alone is not enough; we need to annotate it with genes and other functional elements for the genome to be of greatest use. Knowledge of gene loci is critical to constructing phylogenetic relationships between organisms, and to studying the functional implications of variants, both common uses of reference genomes. While model-based, purely computational gene predictors can be highly accurate in bacteria, gene sparsity and intronic regions make this task more difficult in eukaryotes (Salzberg, 2019). For improved annotations, some RNA-seq information is required (Salzberg, 2019).

Here, as part of our newly developed Methods in Nucleic Acid Sequencing university course, we used a hybrid approach, applying long-read nanopore

sequencing to assemble a highly contiguous genome of *C. nivariensis*, followed by short-read sequencing to polish or correct errors in our assembly. We followed this by a combination of nanopore direct RNA sequencing as well as short-read RNA-seq to annotate our assembly. By combining this data with liftover of annotations from evolutionary "cousins" of *nivariensis*, we have generated a new and annotated reference genome for the community.

3.3 Results

3.3.1 Genome statistics

Using our nanopore and Illumina sequencing data, we generated a new assembly of *Candida nivariensis*, JHU_Cniv_v1 (Methods). Our assembly consists of 11.8 Mb of sequence in 16 contigs with an N50 of 886 Kb (**Figure 3.1**, **Table 4.7**). Compared to the reference genome, we have 275kb of additional sequence, 218kb of which is accounted for by gaps in the reference which are newly spanned by JHU_Cniv_v1. Of the 69 newly spanned gap sequences, 54 were identified as repeat regions. Another 13 gap regions were identified to contain a higher than average proportion of multi-mapping short reads (>10% in gap regions vs 7% average across the genome).

	Contigs	N50	Longest Contig	Shortest Contig	Total Length
Reference	123	248 Kb	807 Kb	666 bp	11.56 Mb
JHU_Cniv_v1	16	886 Kb	1.42 Mb	28.5 Kb	11.83 Mb

Table 3.1: Assembly Statistics. Assembly statistics of JHU_Cniv_v1 and the reference genome for *C. nivariensis*.

To determine whether JHU_Cniv_v1 contigs represent full chromosomes,

we looked for telomere repeats in our assembly and attempted to use related yeast reference genomes to scaffold. In our assembly, 11 contigs terminate at both ends in repeats of CTGGGTGCTGTGGGGT, the telomere sequence of *Candida glabrata* (McEachern and Blackburn, 1994). The other

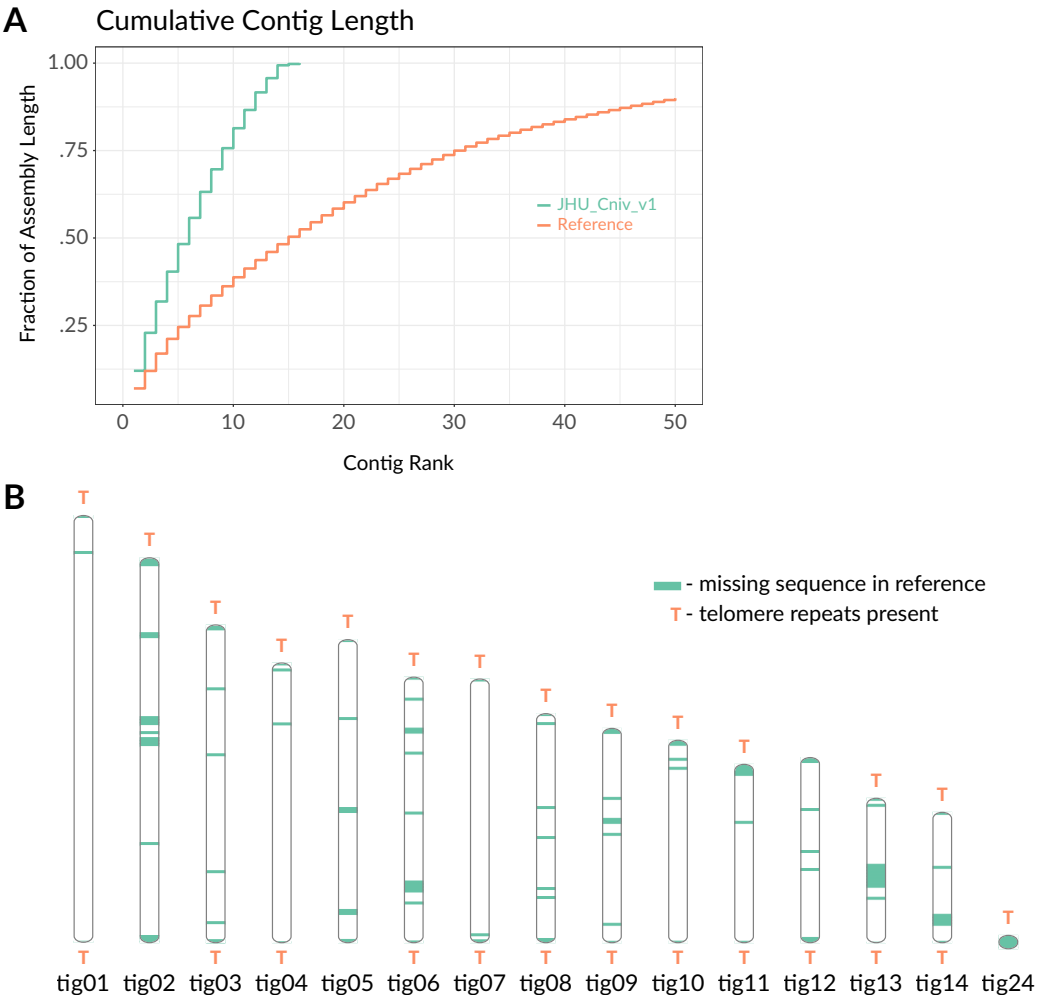

Figure 3.1: Characteristics of the JHU_Cniv_v1 assembly. (A) Cumulative lengths of the 50 longest sequences in our assembly and previous reference genome. **(B)** Ideogram of assembly. Sequence that is missing in the reference genome is shown along each non-mitochondrial contig, and the positions of telomere repeats are marked.

4 non-mitochondrial sequences terminate only at one end in this telomeric repeat (**Figure** 3.1, **Table** 3.2), suggesting they may scaffold to form two additional chromosomes. This suggests that, like *C. glabrata*, the *C. nivariensis* genome also contains 13 chromosomes.

Contig	Length (bp)	Forward Telomeres	Reverse Telomeres
tig01	1423475	35	38
tig02	1283968	0	39
tig03	1060011	35	39
tig04	933062	36	26
tig05	1010854	0	36
tig06	885783	35	38
tig07	879540	39	35
tig08	763992	34	33
tig09	714796	35	47
tig10	675194	36	36
tig11	594828	32	26
tig12	617546	36	0
tig13	481613	38	41
tig14	434809	33	33
tig24	44616	0	39
JHU_Cniv_v1_mito	28512	0	0

Table 3.2: Contig and telomere lengths. Contig lengths and the number of times the forward and reverse telomere sequence appears in each

We tried to further scaffold our assembly using the more contiguous and highly related *glabrata* genome as a reference, but we found that reference

based scaffolders such as Medusa v1.6 (Bosi et al., 2015) and RagTag v1.0.2 (Alonge et al., 2019) either placed telomeric sequences in the middle of scaffolds or made no improvement (**Figure 3.2**). Upon aligning the *C. glabrata* genome to JHU_Cniv_v1 using Mummer, we found only sporadic shared segments of negligible length (**Figure 3.3**), as opposed to a nearly perfect 1:1 alignment between JHU_Cniv_v1 and the current *C. nivariensis* reference genome (**Figure 3.4**). This indicated that the *C. glabrata* genome is not sufficiently similar to *C. nivariensis* to use as a reference for contig scaffolding. Using the *C. nivariensis* reference genome for scaffolding similarly results in erroneous placement of telomere repeats in the middle of scaffolds, or no change to our assembly. This is unsurprising, as the *C. nivariensis* reference genome is so highly fragmented.

3.3.2 Genome completeness

To assess assembly completeness, fungal single-copy orthologs were checked using BUSCO v5.0.0 (Simão et al., 2015) and its available saccharomycetes_odb10 database. Out of 2137 BUSCOs searched, JHU_Cniv_v1 has only 14 missing, 13 of which are also missing in the current reference (**Figure 3.5**). This additional missing gene, RNA polymerase archaeal subunit P/eukaryotic subunit RPABC4 (buscoID 41996at4891), though present in the reference, has the second lowest combined match length and match score among all genes searched. From the reference, we extracted the nucleotide sequence of this match using the coordinates reported by BUSCO, and searched for it in JHU_Cniv_v1 using BLAST. We found a full-length match with 99.9% identity, suggesting that this

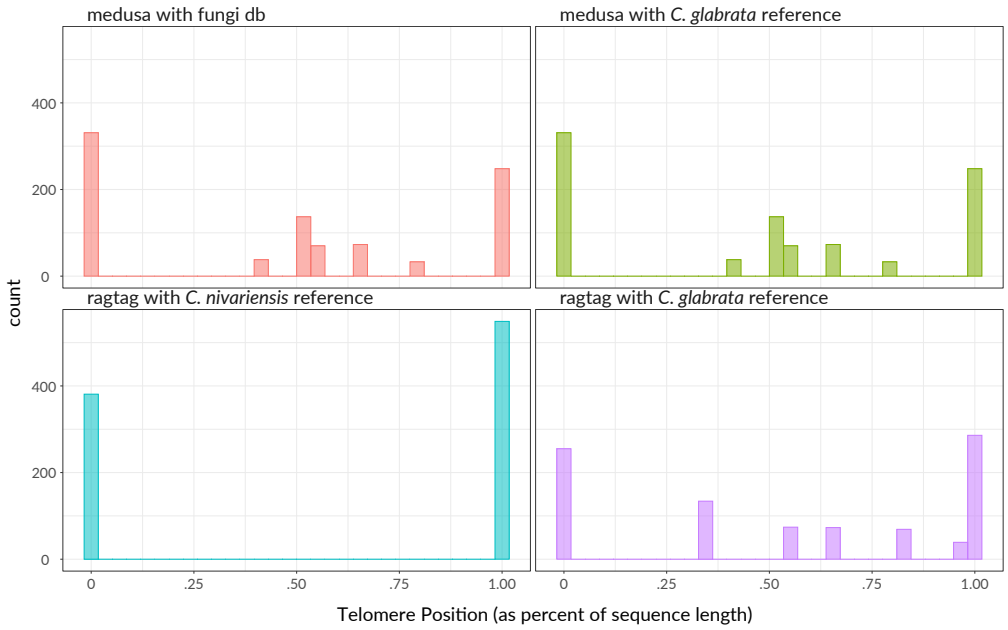

Figure 3.2: Telomere positions reference based scaffolds. Histogram of telomere repeat positions in our assembly, and in scaffolds produced by RagTag and MeDuSa. When MeDuSa is used with a database including the reference genomes of *C. nivariensis*, *C. glabrata*, *C. bracarensis*, and *N. delphensis*, telomeres are placed in the middle of contigs. The same result is produced when only the *C. glabrata* genome is used for scaffolding with MeDuSa, and MeDuSa fails to run when only the *C. nivariensis* reference is used. When the *C. nivariensis* reference genome is used for scaffolding with RagTag, no changes are made. When the more contiguous *C. glabrata* genome is used with RagTag, telomere sequences are again placed in the middle of sequences, suggesting a scaffolding error.

BUSCO is not actually absent in JHU_Cniv_v1. Upon further examination of this alignment, we found that all seven nonmatching nucleotides consist of small deletions associated with poly-A or poly-T homopolymers, known error-prone regions for nanopore sequencing data (Watson and Warr, 2019).

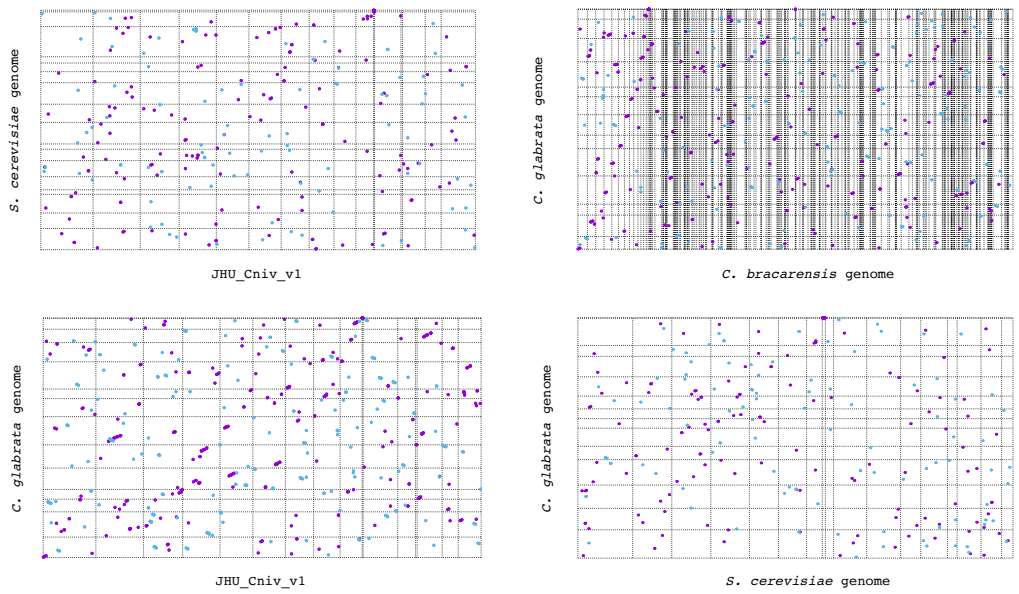

Figure 3.3: Whole genome alignments between related yeasts. Whole genome alignment of our new assembly against the *S. cerevisiae* (top left), and *C. glabrata* (bottom left) reference genomes. For both, there are no long alignments, suggesting that there is little similarity in genome structure between these species and *C. nivariensis*. *C. bracarensis*, a close relative to both *C. glabrata* and *C. nivariensis*, also shares little genome similarity to *C. glabrata* (top right), suggesting that yeast genomes within the *glabrata* clade are not generally similar enough to support inter-species reference based scaffolding. We also compared *C. glabrata* to the highly contiguous and complete *S. cerevisiae* genome (bottom right) to check that genome contiguity alone did not bias the genome similarity detected.

3.3.3 Repetitive genes

As *C. glabrata* subtelomeric regions have been proven to be difficult to correctly assemble using short-read data (Xu et al., 2020), we compare the copy number of *C. glabrata* subtelomere gene homologs between the *C. nivariensis* reference genome and JHU_Cniv_v1. Using the assembly and re-annotation of *C. glabrata* from Xu et al. (2020), we extracted the sequences of the *C. glabrata* subtelomere genes and used BLAST (v2.6.0+) to find any matches in the *C.*

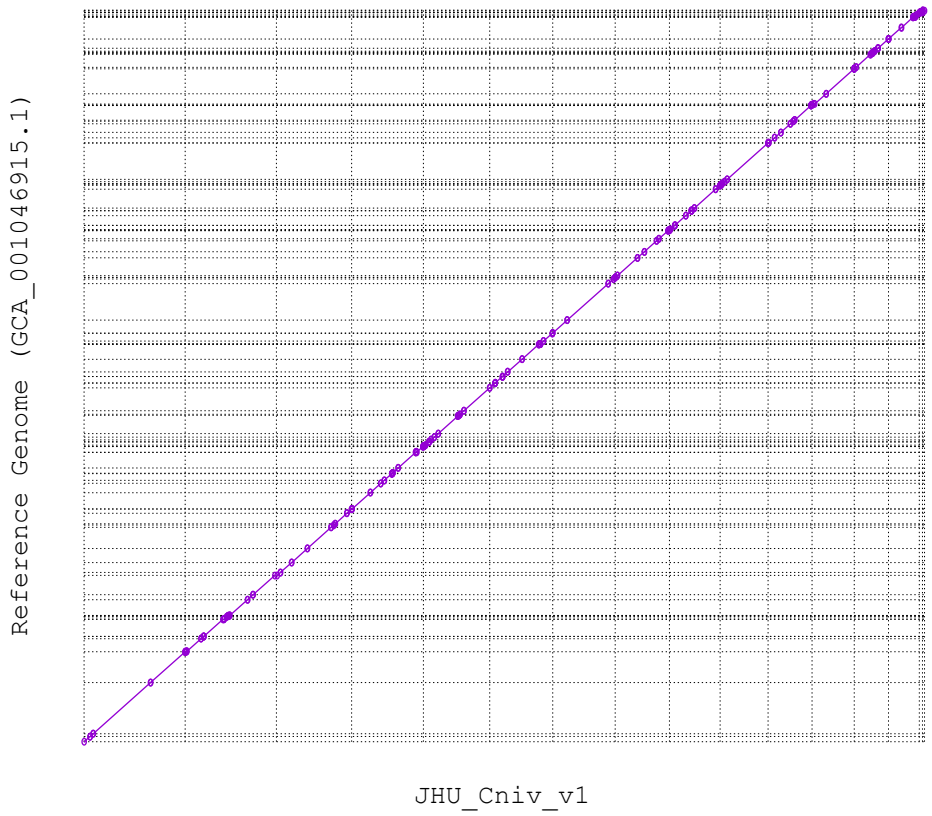

Figure 3.4: Whole genome alignment of JHU_Cniv_v1 and the *C. nivariensis* reference genome. Whole genome alignment of the current reference genome (y axis) compared to our new assembly (x axis). Alignments match with no notable structural variants, and very little missing or duplicated sequence.

nivariensis reference and JHU_Cniv_v1. We observed an identical set of 48 *C. glabrata* subtelomere genes in both *C. nivariensis* genomes but found that the copy number for several genes was greater in JHU_Cniv_v1 (**Figure** 3.6). To account for genes truncated by short contigs in the reference genome, we calculate copy number by summing the alignment lengths of all the hits of a particular gene and dividing by gene length. Of the 48 *C. glabrata* genes with

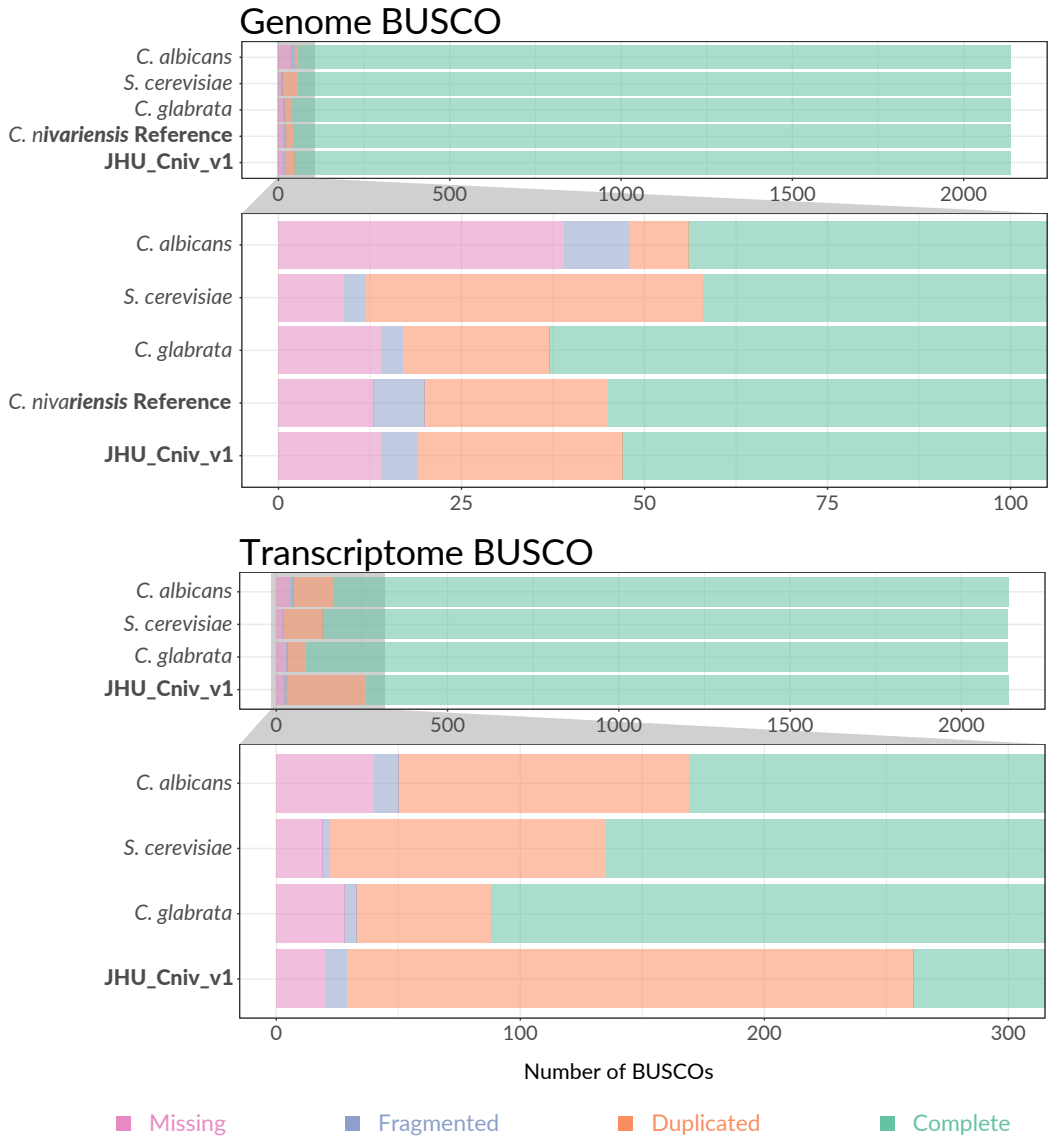

Figure 3.5: Completeness of the JHU_Cniv_v1 assembly. Genome and transcriptome completeness Bar charts comparing BUSCOs detected in JHU_Cniv_v1 and accompanying transcriptome to those of the current *C. albicans*, *S. cerevisiae*, *C. glabrata*, and *C. nivariensis* reference genomes. No reference transcriptome is currently available for *C. nivariensis*.

homology in *C. nivariensis*, 35 are ribosomal. With the exception of just three ribosomal genes, which occur a similar number of times in both *C. nivariensis*

genomes, all homologous ribosomal genes appear once in the reference, and either four or six times in JHU_Cniv_v1 (**Figure 3.6**).

Using JHU_Cniv_v1, we identified GPI-anchored membrane proteins among annotated genes >1000-nt long. Using GffRead (Pertea and Pertea, 2020), we constructed the amino acid sequences for these genes and excluded any with internal stop codons. We then used PredGPI (Pierleoni, Martelli, and Casadio, 2008) to predict which of these encoded GPI proteins, using an FDR cutoff of <0.0005 (Xu et al., 2020) to find 86 total genes. As GPI-anchored fungal adhesins typically contain tandem repeats (Lipke, 2018; Xu et al., 2020), we further filtered for genes overlapping with tandem repeats as classified by Tandem Repeat Finder and identified 53 of the GPI genes as putative adhesins. As with *C. glabrata*, the putative adhesins typically spanned multiple kilobases (**Figure 3.6**), though we do not find very long (>13 kb) genes in contrast to several *glabrata* GPI-CWPs. To find the corresponding adhesin genes in the *C. nivariensis* reference genome, we again used BLAST, and compared the longest hit of each adhesin gene to the true length of the gene as predicted in JHU_Cniv_v1 (**Figure 3.6**). Notably, no hit in the reference genome exceeded 3.5kb, and 27 of these adhesin genes are not found continuously, suggesting the previous reference either truncated or did not continuously assemble these important pathogenicity genes.

3.4 Discussion

JHU_Cniv_v1 is a high quality, extremely contiguous assembly of *Candida nivariensis* constructed by long reads and polished by short reads. It spans

large, repetitive gaps in the *nivariensis* genome that have fragmented short-read assemblies thus far, and includes a full mitochondrial chromosome, as well as telomere repeats. These telomere repeats are identical to those in *C. glabrata* and have been found to be shared within the entire "*glabrata* group" (Gabaldón et al., 2013). The orientation of the telomeres suggests that *C. nivariensis* has 13 chromosomes, which is in agreement with previous pulsed-field gel electrophoresis (PFGE) data (Gabaldón et al., 2013). Furthermore,

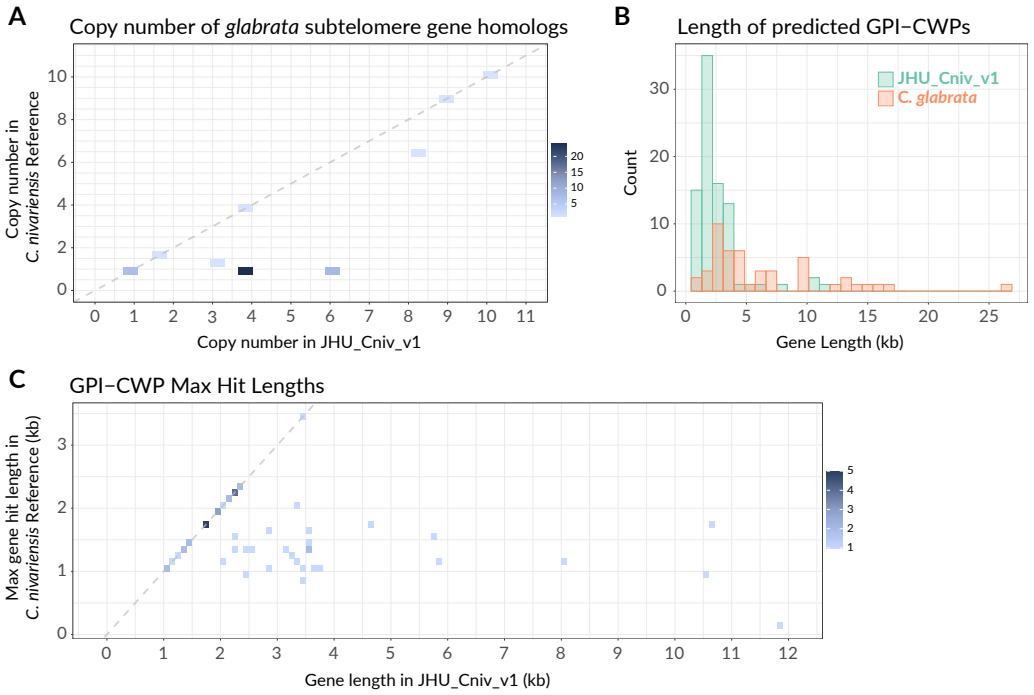

Figure 3.6: GPI genes. (A) Scatterplot showing the number of times each glabrata subtelomere gene homolog appears in the *C. nivariensis* reference genome and in JHU_Cniv_v1. Overlapping points are shown on the color scale, and the y=x line is shown in dashed gray. **(B)** Histogram of adhesion protein lengths in glabrata as annotated by Xu et al., and the lengths of predicted adhesion proteins found in JHU_Cniv_v1. **(C)** Scatterplot showing the maximum BLAST alignment lengths for each predicted nivariensis GPI gene in JHU_Cniv_v1 and the *C. nivariensis* reference genome. Overlapping points are shown on the color scale, and the y=x line is shown in dashed gray.

of the contigs missing telomere repeats on one end, we note that scaffolding tig05 with tig12 and tig02 with tig24 would result in 13 chromosomes that would all match PFGE length estimates to 8% error or less, which is within the expected range of PFGE error for very large DNA fragments (Cutting et al., 1988).

As assessed by BUSCO, genome completeness of the current *C. nivariensis* reference and JHU_Cniv_v1 are comparable to other related yeasts, with our genome slightly improved over the previous reference. However, while JHU_Cniv_v1 is a much more contiguous assembly than any *C. nivariensis* genome preceding it, the few remaining sequence errors still can pose a problem to downstream analyses, as evidenced by the seemingly absent BUSCO we manually identified.

Our accompanying RNA-seq data enabled us to annotate this genome, achieving a similar level of BUSCO completeness to some of the most highly studied model organisms. Our annotation has comparable or lower levels of missing and fragmented BUSCOs compared to the reference annotations, though more duplicated ones. While our annotation is largely comparable to those of similar yeasts (**Table 3.3**), it has not been manually curated, and should thus be treated as preliminary. Of course, as these organisms were grown under only one condition before RNA extraction, it remains unlikely that this annotation is fully complete.

To demonstrate the utility of genome and annotation contiguity, we examine genes from a difficult to assemble region in *C. glabrata*. For each subtelomeric *C. glabrata* gene with homology in *C. nivariensis*, more copies were found

	Total Exons	Total Genes
JHU_Cniv_v1	7,298	5,859
C. glabrata	5,629	5,448
S. cerevisiae	6,760	6,420
C. albicans	6,732	6,263

Table 3.3: Gene and exon counts of JHU_Cniv_v1 and related yeasts. Gene and exon counts of our annotation and currently available reference annotations

in JHU_Cniv_v1, as its contiguity allows it to more easily capture repeated genome elements. We note that of subtelomeric *glabrata* genes found, the majority are ribosomal, and of these, only three do not show a four or six times increased copy number in JHU_Cniv_v1. Due to the repetitive nature of rDNA arrays, it can be difficult for short-read genome assemblies to capture them in their full complexity. Conversely, our long-read assembly more easily spans these regions, potentially providing a clearer look at the biology in which they are involved.

In addition to genes arranged in complex and repetitive patterns, our more contiguous assembly enables analysis of large genes with internal repeats, such as GPI adhesins. Since these genes are so large, it can be difficult or impossible to predict them from fragmented assemblies which are unable to capture them in their full length. As adhesins are critical to understanding elements of pathogenicity in these yeasts, fragmented genome assemblies and missing gene annotations can be crippling to this dimension of research in these organisms.

3.5 Methods

3.5.1 Media and growth conditions

For genomic extractions, a single colony of *C. nivariensis* CBS9983, originally isolated from a blood culture of a Spanish woman (Alcoba-Flórez et al., 2005), was inoculated into synthetic complete (SC) medium supplemented with 2% glucose and shaken overnight at 30°C in a glass culture tube. For RNA extractions, *C. nivariensis* CBS9983 was grown to log phase in SC medium supplemented with 2% glucose at 30°C in a glass culture tube.

3.5.2 DNA isolation and sequencing

DNA was extracted from liquid culture using the Zymo Fungal/Bacterial DNA MiniPrep Kit according to manufacturer specifications. Two ONT sequencing libraries were prepared from the extracted DNA using the ONT rapid barcoding sequencing kit (SQK-RBK004), and each was sequenced on a separate MinION flowcell (R9.4). Two Illumina libraries were prepared with the Nextera Flex Library Prep Kit, each using 400ng of extracted DNA. Both Illumina libraries were then sequenced on a single iSeq 100 run.

3.5.3 RNA isolation and sequencing

RNA was extracted from liquid culture using the Zymo Fungal/Bacterial RNA MiniPrep Kit. Using the NEBNext Poly(A) mRNA Magnetic Isolation Module, polyA tailed mRNA was isolated from the total RNA. Two ONT direct RNA sequencing libraries were prepared and sequenced on separate MinION

flowcells, each using 200ng of polyA selected RNA and the SQK-RNA002 sequencing kit. With the NEBNext Ultra II RNA First-Strand Synthesis Module and the NEBNext Ultra II Non-Directional RNA Second Strand Synthesis Module, cDNA was prepared from the isolated mRNA. Two individual Illumina libraries were then prepared with the Nextera Flex Library Prep Kit, each using 400ng of cDNA. Both library replicates were then sequenced on a single iSeq 100 run, generating 2 × 150 paired-end reads.

3.5.4 Genome assembly

Nanopore data were basecalled using Guppy v3.2.4 on default settings. Reads greater than 3kb long with an average basecalling quality score greater than 7 were assembled into 21 contigs using Canu v2.1 (Koren et al., 2017) on default settings with the genome size set to 11m. Illumina DNA reads were trimmed for adapters and quality using Trimmomatic v0.39 (Bolger, Lohse, and Usadel, 2014) using settings `LEADING:3 TRAILING:3 SLIDINGWINDOW:4:30 MINLEN:36`. The trimmed reads were then used to iteratively correct draft assembly using Freebayes v1.3.4-pre1 (Garrison and Marth, 2012) with alignments made by bwa mem v0.7.17-r1198-dirty (Li, 2013) using default settings. Changes were made at positions where both the alternative allele frequency was greater than 0.5 and the total number of alternate allele observations was greater than 5. We aligned and corrected the assembly iteratively for three rounds, after which further rounds of corrections made no changes.

Of our 21 corrected contigs, 5 were flagged as repeats by Canu and originally constructed from fewer than 180 nanopore reads. The remaining 16

contigs were constructed from over 1800 nanopore reads each. Because the five repetitive contigs were constructed from so few reads and were found to occur elsewhere in the assembly through Mummer v4.0.0rc1 (Marçais et al., 2018) and nanopore read alignment Minimap2 v2.17 (Li, 2018), we excluded them from the final assembly. One 32-Kb contig was suggested to be circular by Canu, and therefore likely to be a mitochondrial sequence. To confirm, we aligned this contig to the complete mitochondrial genome of *C. nivariensis* (NCBI: NC_036379.1) using Mummer, and observed a 3662-bp sequence in the reference mitochondrial genome which appears at both ends of our 32-kb circular contig. Using the Mummer alignments (**Figure** 3.7), we removed the extraneous 3662bp from the end of our contig, resulting in a 28-kb mitochondrial genome, which we named "JHU_Cniv_v1_mito." Lastly, we remapped the ONT and Illumina reads back to the assembly, and found no bases with zero coverage, indicating that none of our contigs need to be further broken (**Figure** 3.8). Henceforth, we refer to this assembly as "JHU_Cniv_v1."

Repeat regions were identified by Tandem Repeats Finder v4.09 (Benson, 1999) with settings (Xu et al., 2020): `match = 2, mismatch = 7, delta = 7, pm = 80, pi = 10, minscore = 50, maxperiod = 600`. Multimapping short reads were identified using bwa mem (Li, 2013) on default settings.

3.5.5 Annotation

Illumina RNA-seq reads were trimmed using Trimmomatic v0.39 (Bolger, Lohse, and Usadel, 2014) in order to check for any remaining adapter sequences and to filter out reads with low base quality. HISAT2 v2.1.0 was

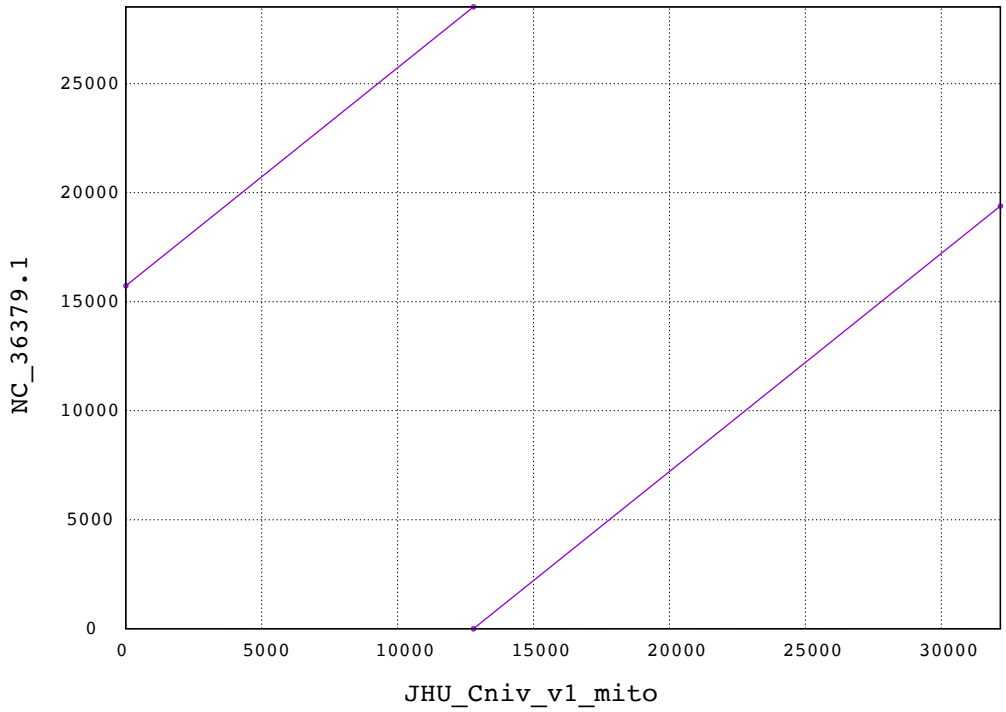

Figure 3.7: Alignment of JHU_Cniv_v1 mitochrondrial contig and the *C. nivariensis* mitochondrial genome. Alignment of our 32Kb circular contig (x axis) with the completed mitochondrial genome of the *C. nivariensis* reference genome (y axis). The final 3662bp of this contig appears twice in the reference genome.

used on default settings to align the trimmed cDNA reads to the assembly. The BRAKER v2.1.5 pipeline (Hoff et al., 2019) was then used to make gene predictions using these alignments. Currently, ONT dRNA compatibility with BRAKER is in development, and that data was thus not used for prediction. Instead, ONT dRNA reads were aligned to the genome assembly using Minimap2 on recommended settings for nanopore direct RNA reads (`-ax splice -uf -k14`). Transcripts were then assembled from the dRNA alignments using StringTie2 v2.1.5 (Kovaka et al., 2019) with the long read option (`-L`). Using Liftoff v1.5.0 (Shumate and Salzberg, 2020), we lifted over the annotations

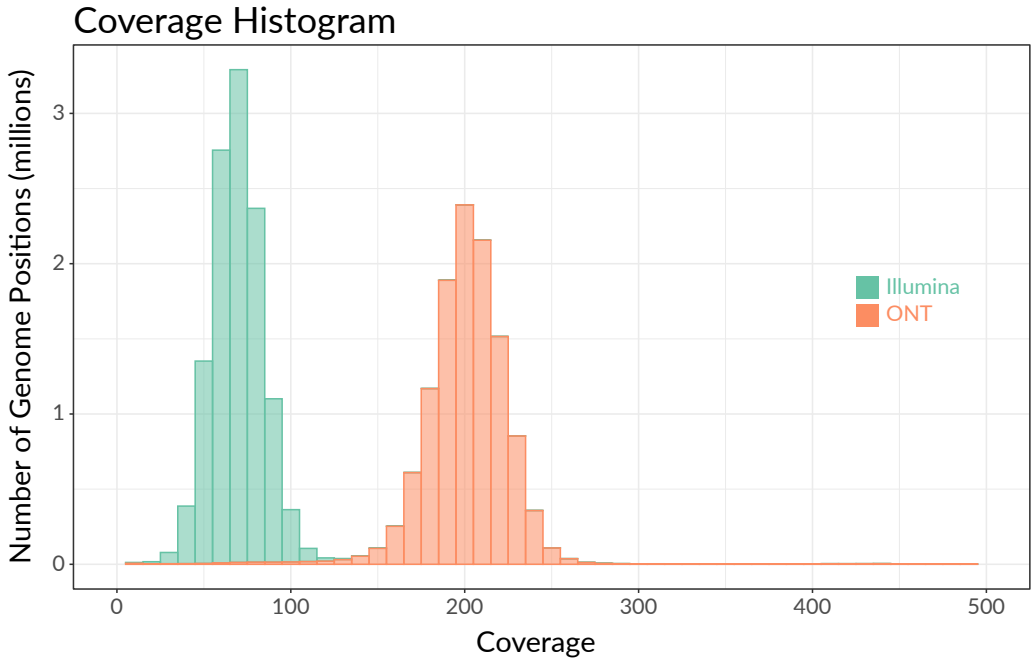

Figure 3.8: Coverage histograms. Histogram of coverage per base in our assembly by filtered (>3kb) ONT reads and trimmed Illumina reads.

from *C. glabrata* (NCBI: GCF_000002545.3), *Saccharomyces cerevisiae* (NCBI: GCF_000146045.2), *Candida albicans* (NCBI: GCF_000182965.3).

Starting with the BRAKER predictions, GffCompare v0.12.1 (Pertea and Pertea, 2020) was used to add nonoverlapping annotations lifted from *C. glabrata*, *S. cerevisiae*, and *C. albicans* in that order. Specifically, we add any annotation with class code "u" in the GffCompare .tmap outputs when comparing our list of genes with a list of potential genes to add, since these refer to intergenic regions devoid of any overlap or proximity to previous annotations. Finally, we compared and added nonredundant transcripts assembled by StringTie2 to the annotation using GffCompare.

	Total	Gene	Exon
Augustus (BRAKER)	23,497	5,028	6,109
Genemark.hmm (BRAKER)	36	6	12
Liftoff *glabrata*	263	130	2
Liftoff *cerevisiae*	42	21	0
Liftoff *albicans*	0	0	0
StringTie	2,141	824	1,175

Table 3.4: Contributions from each annotation software. Number of genes and exons added by each software

3.5.6 Data Availability

All sequence data are available in the Sequence Read Archive, under BioProject PRJNA686979. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAEVGP000000000. The version described in this here is version JAEVGP010000000. Code used for analysis is available at https://github.com/timplab/nivar.

References

Borman, Andrew M, Rebecca Petch, Christopher J Linton, Michael D Palmer, Paul D Bridge, and Elizabeth M Johnson (2008). "*Candida nivariensis*, an emerging pathogenic fungus with multidrug resistance to antifungal agents". en. In: *J. Clin. Microbiol.* 46.3, pp. 933–938.

Aznar-Marin, Pilar, Fátima Galan-Sanchez, Pilar Marin-Casanova, Pedro García-Martos, and Manuel Rodríguez-Iglesias (2016). "*Candida nivariensis* as a New Emergent Agent of Vulvovaginal Candidiasis: Description of Cases and Review of Published Studies". en. In: *Mycopathologia* 181.5-6, pp. 445–449.

Gabaldón, Toni, Tiphaine Martin, Marina Marcet-Houben, Pascal Durrens, Monique Bolotin-Fukuhara, Olivier Lespinet, Sylvie Arnaise, Stéphanie Boisnard, Gabriela Aguileta, Ralitsa Atanasova, Christiane Bouchier, Arnaud Couloux, Sophie Creno, Jose Almeida Cruz, Hugo Devillers, Adela Enache-Angoulvant, Juliette Guitard, Laure Jaouen, Laurence Ma, Christian Marck, Cécile Neuvéglise, Eric Pelletier, Amélie Pinard, Julie Poulain, Julien Recoquillay, Eric Westhof, Patrick Wincker, Bernard Dujon, Christophe Hennequin, and Cécile Fairhead (2013). "Comparative genomics of emerging pathogens in the *Candida glabrata* clade". en. In: *BMC Genomics* 14, p. 623.

Croll, Daniel, Marcello Zala, and Bruce A McDonald (2013). "Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen". en. In: *PLoS Genet.* 9.6, e1003567.

Ford, Christopher B, Jason M Funt, Darren Abbey, Luca Issi, Candace Guiducci, Diego A Martinez, Toni Delorey, Bi Yu Li, Theodore C White, Christina Cuomo, Reeta P Rao, Judith Berman, Dawn A Thompson, and Aviv Regev (2015). "The evolution of drug resistance in clinical isolates of *Candida albicans*". en. In: *Elife* 4, e00662.

López-Fuentes, Eunice, Guadalupe Gutiérrez-Escobedo, Bea Timmermans, Patrick Van Dijck, Alejandro De Las Peñas, and Irene Castaño (2018). "*Candida glabrata*'s Genome Plasticity Confers a Unique Pattern of Expressed Cell Wall Proteins". en. In: *J Fungi (Basel)* 4.2.

Carreté, Laia, Ewa Ksiezopolska, Emilia Gómez-Molero, Adela Angoulvant, Oliver Bader, Cécile Fairhead, and Toni Gabaldón (2019). "Genome Comparisons of *Candida glabrata* Serial Clinical Isolates Reveal Patterns of Genetic Variation in Infecting Clonal Populations". en. In: *Front. Microbiol.* 10, p. 112.

Todd, Robert T, Tyler D Wikoff, Anja Forche, and Anna Selmecki (2019). "Genome plasticity in *Candida albicans* is driven by long repeat sequences". en. In: *Elife* 8.

Barry, J D, M L Ginger, P Burton, and R McCulloch (2003). "Why are parasite contingency genes often associated with telomeres?" en. In: *Int. J. Parasitol.* 33.1, pp. 29–45.

De Las Peñas, Alejandro, Shih-Jung Pan, Irene Castaño, Jonathan Alder, Robert Cregg, and Brendan P Cormack (2003). "Virulence-related surface glycoproteins in the yeast pathogen *Candida glabrata* are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing". en. In: *Genes Dev.* 17.18, pp. 2245–2258.

Naumov, G I, E S Naumova, and E J Louis (1995). "Genetic mapping of the alpha-galactosidase MEL gene family on right and left telomeres of *Saccharomyces cerevisiae*". en. In: *Yeast* 11.5, pp. 481–483.

Iraqui, Ismail, Susana Garcia-Sanchez, Sylvie Aubert, Françoise Dromer, Jean-Marc Ghigo, Christophe d'Enfert, and Guilhem Janbon (2005). "The Yak1p kinase controls expression of adhesins and biofilm formation in *Candida glabrata* in a Sir4p-dependent pathway". en. In: *Mol. Microbiol.* 55.4, pp. 1259–1271.

Carreto, Laura, Maria F Eiriz, Ana C Gomes, Patrícia M Pereira, Dorit Schuller, and Manuel A S Santos (2008). "Comparative genomics of wild type yeast strains unveils important genome diversity". en. In: *BMC Genomics* 9, p. 524.

Brown, Chris A, Andrew W Murray, and Kevin J Verstrepen (2010). "Rapid expansion and functional divergence of subtelomeric gene families in yeasts". en. In: *Curr. Biol.* 20.10, pp. 895–903.

Anderson, Matthew Z, Lauren J Wigen, Laura S Burrack, and Judith Berman (2015). "Real-Time Evolution of a Subtelomeric Gene Family in *Candida albicans*". en. In: *Genetics* 200.3, pp. 907–919.

Timmermans, Bea, Alejandro De Las Peñas, Irene Castaño, and Patrick Van Dijck (2018). "Adhesins in *Candida glabrata*". en. In: *J Fungi (Basel)* 4.2.

McCall, Andrew D, Ruvini U Pathirana, Aditi Prabhakar, Paul J Cullen, and Mira Edgerton (2019). "*Candida albicans* biofilm development is governed by cooperative attachment and adhesion maintenance proteins". en. In: *NPJ Biofilms Microbiomes* 5.1, p. 21.

Carreté, Laia, Ewa Ksiezopolska, Cinta Pegueroles, Emilia Gómez-Molero, Ester Saus, Susana Iraola-Guzmán, Damian Loska, Oliver Bader, Cecile Fairhead, and Toni Gabaldón (2018). "Patterns of Genomic Variation in the Opportunistic Pathogen *Candida glabrata* Suggest the Existence of Mating and a Secondary Association with Humans". en. In: *Curr. Biol.* 28.1, 15–27.e7.

Wick, Ryan R, Louise M Judd, and Kathryn E Holt (2019). "Performance of neural network basecalling tools for Oxford Nanopore sequencing". en. In: *Genome Biol.* 20.1, p. 129.

Fox, Edward J, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb (2014). "Accuracy of Next Generation Sequencing Platforms". en. In: *Next Gener Seq Appl* 1.

Watson, Mick and Amanda Warr (2019). "Errors in long-read assemblies can critically affect protein prediction". en. In: *Nat. Biotechnol.* 37.2, pp. 124–126.

Garrison, Erik and Gabor Marth (2012). "Haplotype-based variant detection from short-read sequencing". In: arXiv: 1207.3907 [q-bio.GN].

Walker, Bruce J, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, and Ashlee M Earl (2014). "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement". en. In: *PLoS One* 9.11, e112963.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić (2017). "Fast and accurate de novo genome assembly from long uncorrected reads". en. In: *Genome Res.* 27.5, pp. 737–746.

Salzberg, Steven L (2019). "Next-generation genome annotation: we still struggle to get it right". en. In: *Genome Biol.* 20.1, p. 92.

McEachern, M J and E H Blackburn (1994). "A conserved sequence motif within the exceptionally diverse telomeric sequences of budding yeasts". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 91.8, pp. 3453–3457.

Bosi, Emanuele, Beatrice Donati, Marco Galardini, Sara Brunetti, Marie-France Sagot, Pietro Lió, Pierluigi Crescenzi, Renato Fani, and Marco Fondi (2015).

"MeDuSa: a multi-draft based scaffolder". en. In: *Bioinformatics* 31.15, pp. 2443–2451.

Alonge, Michael, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J Sedlazeck, Zachary B Lippman, and Michael C Schatz (2019). "RaGOO: fast and accurate reference-guided scaffolding of draft genomes". en. In: *Genome Biol.* 20.1, p. 224.

Simão, Felipe A, Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, and Evgeny M Zdobnov (2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs". en. In: *Bioinformatics* 31.19, pp. 3210–3212.

Xu, Zhuwei, Brian Green, Nicole Benoit, Michael Schatz, Sarah Wheelan, and Brendan Cormack (2020). "De novo genome assembly of *Candida glabrata* reveals cell wall protein complement and structure of dispersed tandem repeat arrays". en. In: *Mol. Microbiol.* 113.6, pp. 1209–1224.

Pertea, Geo and Mihaela Pertea (2020). "GFF Utilities: GffRead and GffCompare". en. In: *F1000Res.* 9.

Pierleoni, Andrea, Pier Luigi Martelli, and Rita Casadio (2008). "PredGPI: a GPI-anchor predictor". en. In: *BMC Bioinformatics* 9, p. 392.

Lipke, Peter N (2018). "What We Do Not Know about Fungal Cell Adhesion Molecules". en. In: *J Fungi (Basel)* 4.2.

Cutting, G R, S E Antonarakis, H Youssoufian, and H H Kazazian Jr (1988). "Accuracy and limitations of pulsed field gel electrophoresis in sizing partial deletions of the factor VIII gene". en. In: *Mol. Biol. Med.* 5.3, pp. 173–184.

Alcoba-Flórez, Julia, Sebastián Méndez-Alvarez, Josep Cano, Josep Guarro, Eduardo Pérez-Roth, and María del Pilar Arévalo (2005). "Phenotypic and molecular characterization of *Candida nivariensis* sp. nov., a possible new opportunistic fungus". en. In: *J. Clin. Microbiol.* 43.8, pp. 4107–4111.

Koren, Sergey, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy (2017). "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation". en. In: *Genome Res.* 27.5, pp. 722–736.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data". en. In: *Bioinformatics* 30.15, pp. 2114–2120.

Li, Heng (2013). "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: arXiv: 1303.3997 [q-bio.GN].

Marçais, Guillaume, Arthur L Delcher, Adam M Phillippy, Rachel Coston, Steven L Salzberg, and Aleksey Zimin (2018). "MUMmer4: A fast and versatile genome alignment system". en. In: *PLoS Comput. Biol.* 14.1, e1005944.

Li, Heng (2018). "Minimap2: pairwise alignment for nucleotide sequences". en. In: *Bioinformatics* 34.18, pp. 3094–3100.

Benson, G (1999). "Tandem repeats finder: a program to analyze DNA sequences". en. In: *Nucleic Acids Res.* 27.2, pp. 573–580.

Hoff, Katharina J, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke (2019). "Whole-Genome Annotation with BRAKER". en. In: *Methods Mol. Biol.* 1962, pp. 65–95.

Kovaka, Sam, Aleksey V Zimin, Geo M Pertea, Roham Razaghi, Steven L Salzberg, and Mihaela Pertea (2019). "Transcriptome assembly from long-read RNA-seq alignments with StringTie2". en. In: *Genome Biol.* 20.1, p. 278.

Shumate, Alaina and Steven L Salzberg (2020). "Liftoff: an accurate gene annotation mapping tool". en.

Chapter 4

Methylation based plasmid binning using nanopore sequencing

Portions of this chapter originally appeared in:

Fan Y, Bergman Y, Workman R, Simner P, Timp W. Methylation based plasmid binning using nanopore sequencing. Currently unpublished.

4.1 Abstract

Genomic analysis of microbial communities often involves reconstructing the full genomic complement of the community members after shotgun sequencing. However, contigs generated by metagenome assemblers typically only represent fragments of genomes and need to be further grouped together, a process known as 'binning.' Typical metagenomic binning methods rely primarily on signals from differential coverage and kmer spectra, but these can be confounded by mobile genetic elements, e.g. plasmids, since they can replicate independently of the chromosome and horizontally transfer between organisms. Current methods to resolve mobile genetic elements require more

complex sample preparation or matched pairs of sequencing runs, making them more expensive and less accessible. A less used virtue of microbial systems is the "methylation fingerprint" where different bacteria have different methylation signatures as part of their restriction-modification (R-M) system. Using this, we tested the efficacy of metagenomic binning using methylation signal from a single ONT sequencing run. We find that we can correctly bin plasmids in a standard microbial community sample, and that we can identify a single plasmid appearing in different bacterial hosts. Furthermore, we applied this method to a clinical sample, and find that the results are largely consistent with binning methods based on proximity ligation.

4.2 Introduction

Metagenomics is the analysis of DNA extracted from whole microbial communities for resolving and linking the identities and functions of the community members (Strous et al., 2012). Common analysis steps involve assembling raw sequence reads into contigs, and clustering, or 'binning' them into metagenome assembled genomes (MAGs). Ideally, each MAG then represents the complete genetic complement of a community member, including any non-chromosomal DNA. From these MAGs, microbes can then be identified and genomic potential for functions such as metabolism, antimicrobial resistance, and pathogenicity can be inferred.

A number of computational tools have been developed for binning, the vast majority of which rely on k-mer spectra and coverage signals to differentiate between contigs originating from different organisms (Yue et al.,

2020). While these methods work well for binning chromosomal contigs, they struggle to effectively bin plasmids (Beaulaurier et al., 2018). This is because plasmids are mobile elements which replicate independently from the bacterial chromosome, characteristics which confound the two key signals most binning methods rely upon.

Molecular biology methods like proximity ligation (Hi-C) have been used to successfully link MGEs and their hosts' chromosomes (Burton et al., 2014). DNA inside of intact cells is crosslinked, then extracted and fragmented. Crosslinked fragments of DNA are then ligated together, and sequenced on a single read. Reads partially aligning to a chromosome and partially aligning to an MGE then provides evidence that the chromosome and MGE originated in a single cell. While this method has been shown to work well, it's still somewhat limited by its need for two sequencing runs, intact bacterial samples, and longer, more labor intensive library preparation (Beyi et al., 2021).

More recently, binning methods that use DNA methylation detected by single molecule sequencing technologies have emerged (Beaulaurier et al., 2018; Tourancheau et al., 2021). These methods take advantage of the restriction-methylation systems in bacteria, which cause unique strains of bacteria to exhibit DNA methylation at potentially unique combinations of motifs. Importantly, these methylation patterns remain consistent across all of the chromosomes and plasmids of an individual strain, making them useful for associating any non-contiguous segments of DNA to each other. Methylation based binning with the Pacific Biosciences (Pacbio) single molecule real time

(SMRT) sequencing platform relies on detecting pauses in the raw fluorescence signal (interpulse duration; IPD) to detect methylation. Currently available methods using the Oxford Nanopore Technologies (ONT) platform rely on comparing the raw electrical data generated by ONT sequencers in matched runs of methylated native DNA and completely unmethylated whole genome amplified (WGA) DNA (Tourancheau et al., 2021). Characteristic differences in electrical signal between the two runs can be used to locate and identify methylation. While this also works well, it requires two ONT sequencing runs and additional preparation work (WGA), again increasing the cost and limiting the scope of the method.

Here, we demonstrate a framework to perform metagenomic binning using methylation calls from a single single molecule sequencing run. Recent modifications to the BAM file specification have enabled base modification calls to be encoded from either PacBio or ONT data directly in the BAM file. In our case, we used Megalodon to call methylation, a tool part of ONT's stack of software, in conjunction with the all-context 5mC and 6mA modification model in Rerio, ONT's suite of 'research release' basecalling models. On each metagenomic contig or reference sequence, we determine the methylation level at a combination of motifs, and use similarities in methylation levels at these motifs for binning. To assess binning performance, we examine a shared plasmid in a two-bacteria system, as well as a simple, synthetic microbial community. We then extend the method to a human microbiome sample, and compare to Hi-C results on the same sample.

4.3 Results and Discussion

4.3.1 Microbial Community Standard

To test the feasibility of associating plasmids to chromosomes, we used the ZymoBIOMICS Microbial Community Standard, which contains seven different species of bacteria and one yeast **Table** 4.1. Among the bacteria in this sample, the *E. coli* strain contains a 110kb plasmid, and the *S. aureus* strain contains 3 plasmids with lengths 6.3kb, 2.2kb, and 2.9kb. As the identity of each species in the sample is known, we used their reference genomes to call methylation, and use these methylation signatures to bin plasmids with bacterial chromosomes.

Species	Type II 6mA motifs (listed in McIntyre et al. 2017)	Type II 5mC motifs (WGBS data from McIntyre et al. 2017)
Bacillus subtilis		CmWGG
Escherichia coli	GaTC	CmWGG; TCmGGA; GCmGGC
Enterococcus faecalis	CTKVaG; CTCCaG	
Listeria monocytogenes	ANARaGTANYR	
Pseudomonas aeruginosa		CMTmGAKG
Staphylococcus aureus		GATm
Salmonella enterica	GaTC; CAGaG; BATGCaTV	CmWGG

Table 4.1: Summary of known methylation motifs in the ZymoBIOMICS sample. 5mC modifications are denoted as 'm,' and 6mA motifs are denoted as 'a.'

To select the 'barcode,' or set of methylation motifs that would best differentiate between the species in the sample, we used the 6mA motifs and WGBS data published in McIntyre (McIntyre et al., 2017) et al. From the WGBS data,

we identified the short (<8bp) 5mC motifs methylated in each species, and confirmed that these short motifs account for >99% of all 5mC methylation in these species **Table 4.2**. For both 6mA and 5mC motifs, only the short motifs of Type II and Type III MTases were included, since they comprise 85 of the 100 most frequently occurring methylation specificities listed in REBASE (Roberts et al., 2003). While excluding Type I motifs excludes the potential discriminatory power they provide, it also simplifies analysis by avoiding the long, ambiguous sections of the Type I bipartite motifs, e.g. EcoKI (AAC[N6]GTGC or GCAC[N6]GTT). Due to the absence of many motifs in the short *S. aureus* plasmids, we further narrowed the barcode to just GATC, CAGAG, CCWGG and CTKVAG **Table 4.1**. Despite this motif reduction, the two shortest *S. aureus* plasmids still did not have methylation calls for all four motifs **Table 4.3, Figure 4.1**, and were therefore excluded from further analysis. According to the WGBS and PacBio data (McIntyre et al., 2017) **Tables 4.1, 4.1**, this four-motif barcode is theoretically sufficient to distinguish between all species except *L. monocytogenes* and *P. aeruginosa*, neither of which is methylated at any of the four motifs in the barcode. However, this reduction sufficiently preserves the methylation motifs specific to *S. aureus* and *E. coli* such that they are uniquely distinguishable from the other species in the sample, enabling us to correctly assign these plasmids to their hosts.

For each chromosome and plasmid, we constructed a methylation signature consisting of the average percent methylation at each of the four motifs in the barcode **Figure 4.2a**. The percent methylation at each locus is calculated as the number of methylated calls out of the total number of calls made.

Euclidean distance between these signatures was then used as a measure of similarity **Figure 4.2b**. Notably, the methylation signatures closest to each of the plasmids correspond to the bacterial chromosomes of the species harboring each plasmid respectively. As expected, the distance between *L. monocytogenes*

Chromosome	5mC motif	Number of calls	% methylation with unknown motif
BS.pilon.polished.v3.ST170922	CCAGG	1397	
BS.pilon.polished.v3.ST170922	CCTGG	1492	
BS.pilon.polished.v3.ST170922	unknown	3	0.1
Escherichia_coli_plasmid	CCAGG	140	
Escherichia_coli_plasmid	CCTGG	209	
Escherichia_coli_plasmid	GCCGGC	10	
Escherichia_coli_plasmid	TCCGGA	1	
Escherichia_coli_plasmid	unknown	1	0.28
Escherichia_coli_chromosome	CCAGG	12107	
Escherichia_coli_chromosome	CCTGG	12074	
Escherichia_coli_chromosome	GCCGGC	590	
Escherichia_coli_chromosome	TCCGGA	1689	
Escherichia_coli_chromosome	unknown	1	0
Pseudomonas_aeruginosa_complete_genome 6,792Mb	CATCGAGG	847	
Pseudomonas_aeruginosa_complete_genome 6,792Mb	CATCGATG	506	
Pseudomonas_aeruginosa_complete_genome 6,792Mb	CCTCGAGG	388	
Pseudomonas_aeruginosa_complete_genome 6,792Mb	CCTCGATG	844	
Pseudomonas_aeruginosa_complete_genome 6,792Mb	unknown	1	0.04
Salmonella_enterica_complete_genome 4.760Mb	CCAGG	11465	
Salmonella_enterica_complete_genome 4.760Mb	CCTGG	11325	
Salmonella_enterica_complete_genome 4.760Mb	unknown	4	0.02
Staphylococcus_aureus_chromosome	GATC	9857	
Staphylococcus_aureus_chromosome	unknown	3	0.03
Staphylococcus_aureus_plasmid1	GATC	36	
Staphylococcus_aureus_plasmid1	unknown	0	0
Staphylococcus_aureus_plasmid2	GATC	4	
Staphylococcus_aureus_plasmid2	unknown	0	0
Staphylococcus_aureus_plasmid3	GATC	4	
Staphylococcus_aureus_plasmid3	unknown	0	0

Table 4.2: 5mC methylation motifs in the ZymoBIOMICS sample. For each DNA sequence in the Zymo sample, counts of 5mC motifs are listed, along with the number of methylation loci not attributable to any of the listed motifs.

and *P. aeruginosa* are among the smallest between chromosomes because no motif included in the barcode differentiates between them. However, *E. faecalis* also groups closely with *L. monocytogenes* and *P. aeruginosa*. The only motif included in the barcode which separates *E. faecalis* from *L. monocytogenes* and *P. aeruginosa* is a 6mA motif (CTKV6mAG), and upon examination of the methylation signatures themselves, it becomes clear that this motif does not offer as much discriminatory power as the others **Figure 4.2a**. In fact, removing this

Sequence	Mean Coverage
BS.pilon.polished.v3.ST170922	395.51
Enterococcus_faecalis_complete_genome	477.2
Escherichia_coli_chromosome	272.53
Escherichia_coli_plasmid	184.96
Listeria_monocytogenes_complete_genome	738.13
Pseudomonas_aeruginosa_complete_genome	158.29
Salmonella_enterica_complete_genome	231.21
Staphylococcus_aureus_chromosome	773.49
Staphylococcus_aureus_plasmid1	1590.06
Staphylococcus_aureus_plasmid2	3.4
Staphylococcus_aureus_plasmid3	4.05

Table 4.3: Zymo mean coverage. Mean coverage per sequence in the ZymoBIOMICS sample.

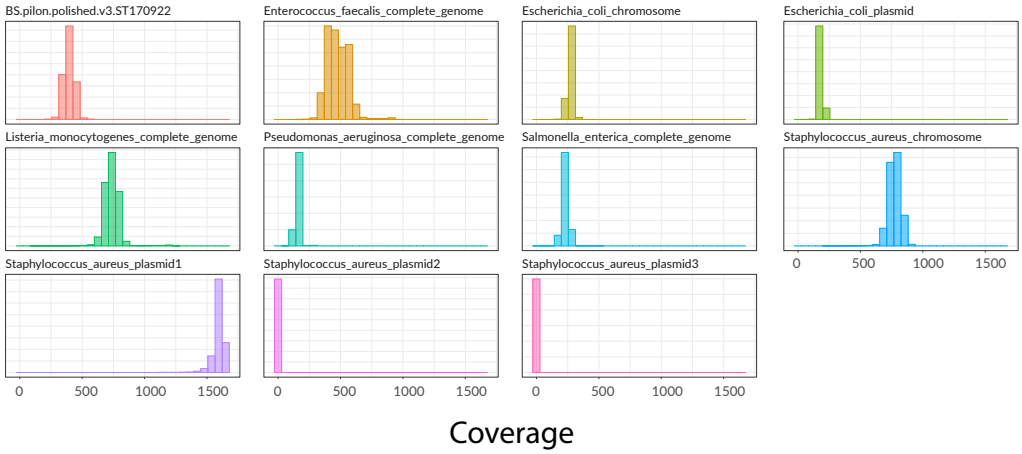

Figure 4.1: Zymo coverage per sequence. Coverage distributions per sequence in the ZymoBIOMICS sample.

motif from the barcode does not appreciably change the distances between the chromosomes **Figure 4.3a**.

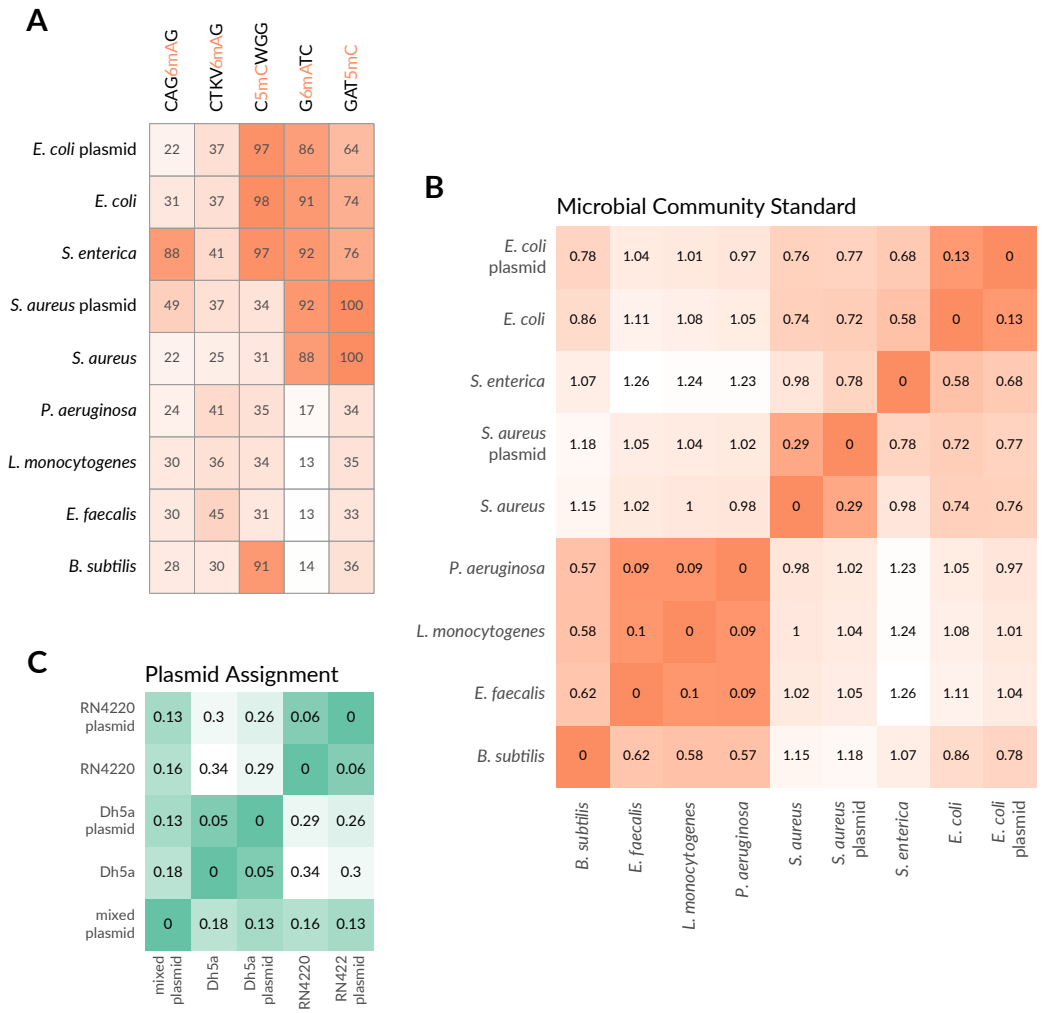

Figure 4.2: Methylation binning in synthetic communities. (A) Percent methylation at each motif for each sequence in the Zymo community. **(B)** Methylation distance between sequences in the Zymo community based on the motifs in **(A)**. **(C)** Methylation distance between RN4220, Dh5a, their associated plasmids, and the synthetic mixture of their plasmids, when percent methylation is calculated using normalized coverage.

To further explore, we implemented a barcode composed of only the

relevant 5mC motifs **Table** 4.1. Due to the lack of CMT(5mC)GAKG occurrences on the *S. aureus* plasmid, we only classified the *E. coli* plasmid using this all-5mC barcode **Figure** 4.3**b**. However, because the newly included

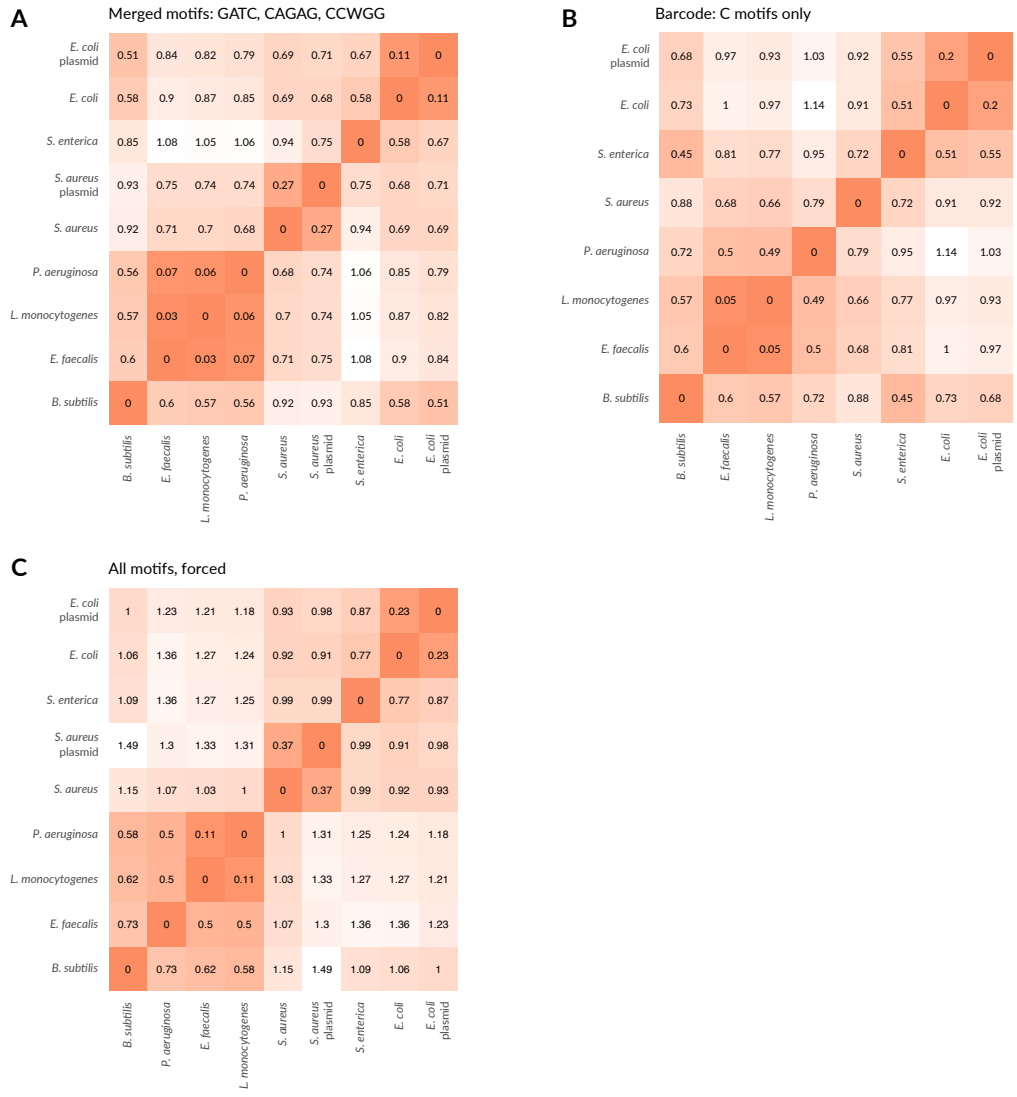

Figure 4.3: Methylation binning in Zymo community. (A) Methylation distance between sequences in the Zymo community, based on the motifs GATC, CAGAG, CCWGG where GATC represents both G6mATC and GAT5mC. **(B)** Methylation distance between sequences in the Zymo community, based on only 5mC motifs GATC, CCWGG, GCCGGC, CMTCGAKG.

CMT5mCGAKG is unique to *P. aeruginosa*, it is differentiable from *L. monocytogenes* and *E. faecalis* using this barcode. As expected, since the *L. monocytogenes* and *E. faecalis* strains in this sample do not exhibit any Type II 5mC methylation, these chromosomes remain close together in methylation space.

Methylation distance also can be calculated with tolerance for missing methylation values, where distances with missing values are scaled to as to be comparable to complete distances. By including all 5mC motifs in the barcode and forcing the classification of the *S. aureus* plasmid, *P. aeruginosa* again becomes differentiable from *L. monocytogenes* and *E. faecalis* **Figure 4.3c**. However, the distance between the *S. aureus* plasmid and chromosome also increases. While the shortest distances observed are still attributable to sequences with identical methylation, these distances are not as tight.

Notably, the methylation distances between the species in this community standard do not closely reflect their taxonomic lineages **Figure 4.4**. This indicates that methylome diversity is not determined by phylogeny, and suggests that methylation can potentially be used to distinguish between even closely related organisms.

4.3.2 Two-bacteria System

We then tested the efficacy of methylation signatures for accurately linking a single plasmid to different bacterial hosts. To do this, we used two bacteria, *E. coli* (strain DH5) and *S. aureus* (strain RN4220), each containing an identical 10kb plasmid (pRW62). This plasmid contains an origin of replication for *E. coli* and a separate origin of replication for *S. aureus* so that it is replicated

in both species. Both were cultured separately and sequenced separately on individual Flongle runs **Figure 4.4**. With bisulfite sequencing data, we observed only two methylated loci in the entire genome of this strain of *S. aureus*, neither of which are in a known 5mC methylation context. In this strain of *E. coli*, all 5mC methylated loci occur in a C5mCWGG context.

Given the results of bisulfite analysis, we selected frequently recorded motifs in REBASE which include an adenine, as well as CCWGG. This resulted in a barcode of GATC, GANTC, CCWGG, CAGAG, CTKVAG, and GTWWAC. Because the methylated residues are not necessarily known for each motif, the base with the highest methylation percentage was chosen to represent each motif locus. Again, the methylation signature of each sequence is represented by the average methylation percentages across all loci for each motif.

Using methylation signatures defined by these five motifs, we are able to

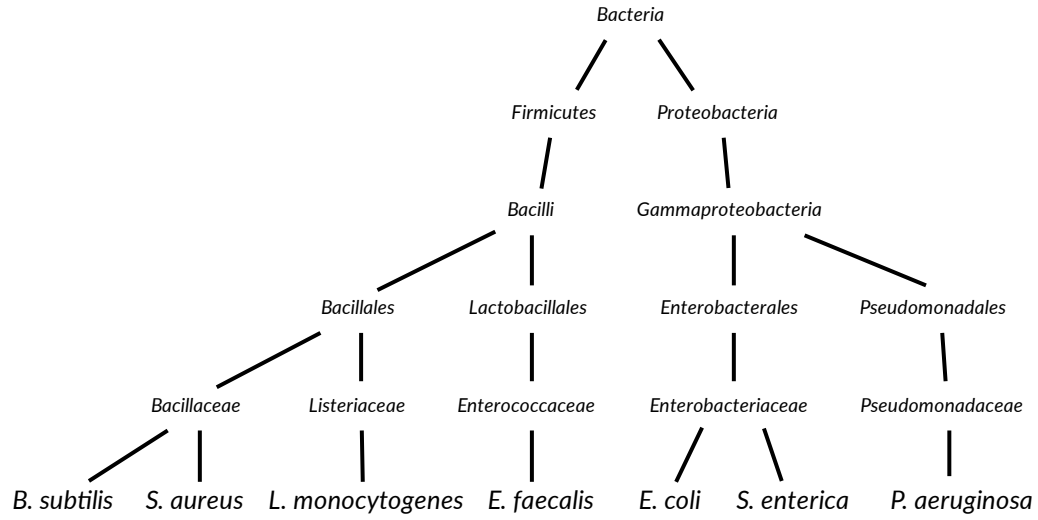

Figure 4.4: Zymo taxonomy. Simplified taxonomic lineages of the Zymo microbial community bacteria.

distinguish the plasmid's presence in *E. coli* from its presence in *S. aureus* **Figure 4.2c**. However, when plasmid reads from both runs are mixed, combined and analyzed, the methylation signature places the mixed plasmid much closer to the *E. coli* signature than the *S. aureus* signature **Figure 4.5a**. This is partially because this plasmid has a much higher copy number in the *E. coli* data **Figure 4.5**. When the *E. coli* plasmid reads are downsampled from 6000X coverage to 440X coverage to match the sequence coverage of the *S. aureus* reads before in silico mixing, the mixed plasmid moves away from the *E. coli* signature **Figure 4.5b**. Much of the remaining bias is accounted for by the high uncertainty of methylation calls at unmethylated loci **Figure 4.6**. Methylation

Run	Yield	Number of reads	Mean read length
Zymo microbial standard ONT	11.16 Gb	668644	16693.9
S. aureus RN4220 ONT	510.67 Mb	222947	2290.54
S. aureus RN4220 Bisulfite-seq read1	78.60 Mb	537075	146.35
S. aureus RN4220 Bisulfite-seq read2	78.76 Mb	537075	146.65
E. coli DH5a ONT	470.85 Mb	177873	2647.11
E. coli DH5a Bisulfite-seq read1	62.53 Mb	430497	145.247
E. coli DH5a Bisulfite-seq read2	62.86 Mb	430497	146.016
Clinical stool ONT	4.71 Gb	735613	6396.76
Clinical stool ONT no human reads	4.65 Gb	691781	6717.72
Clinical stool Hi-C read1	11.22 Gb	112205758	100
Clinical stool Hi-C read2	11.22 Gb	112205758	100
Clinical stool Hi-C read1 no human reads	11.15 Gb	111520031	100
Clinical stool Hi-C read2 no human reads	11.15 Gb	111520031	100
Clinical stool shotgun read1	7.12 Gb	71219590	100
Clinical stool shotgun read2	7.12 Gb	71219590	100
Clinical stool shotgun read1 no human reads	6.85 Gb	68473048	100
Clinical stool shotgun read2 no human reads	6.85 Gb	68473048	100

Table 4.4: Summary statistics of sequencing runs. Yield, number of reads, and mean read length for each sequencing run performed.

callers such as Megalodon or nanopolish compute a probability that a base of a sequencing read is methylated (Simpson et al., 2017). This probability is then thresholded to call bases as either methylated, unmethylated, or uncertain. Though 5-methylcytosine in a CG context gives robust calls of either methylated or unmethylated (Simpson et al., 2017), non-CG 5-methylcytosine models are less robust, and 6-methyladenine gives a smaller electrical signal shift, leading to more uncertain calls. Methylated motifs still present a high probability of methylation, but unmethylated motifs are less clear **Figure 4.6, Table 4.6**. Unmethylated loci are less likely to contribute any methylation calls, suggesting that most calls made on the mixed plasmid originated from the *E.*

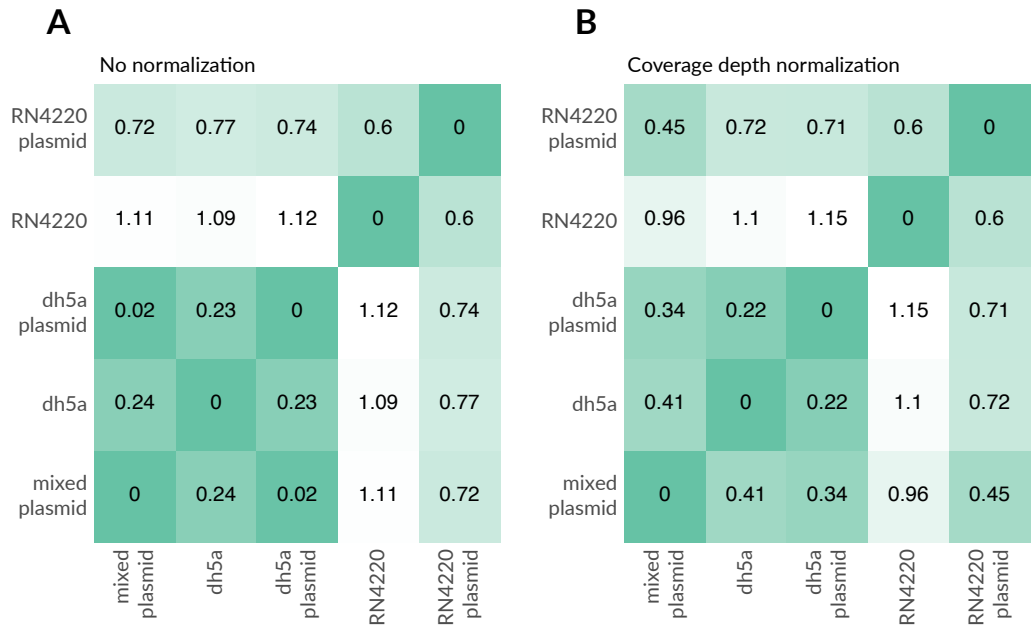

Figure 4.5: Methylation distance between RN4220 and Dh5a. (A) Methylation distance between RN4220 and Dh5a and plasmids with no normalization. The mixed plasmid comprises all plasmid reads. **(A)** Methylation distance between RN4220 and Dh5a where reads have been randomly downsample such that chromosomes have the same average coverage and individual plasmids have the same average coverage. The mixed plasmid contains the downsampled plasmid reads of both strains.

coli data, the only organism with any confirmed methylation. To counteract this, we calculated methylation percentage as the ratio of methylated calls to the total coverage, instead of total methylation calls, at each locus. This way, the unmethylated calls do not play a role in determining the methylation signatures. With this adjustment, the mixed plasmid falls roughly equidistant between the *E. coli* and *S. aureus* signatures **Figure 4.2c**. As methylation callers improve, we expect that this correction may not be necessary, because the values from methylation vs coverage and methylation vs called will converge.

Culture	Sequence	Mean Coverage
S. aureus RN4220 with PRW62	*E. coli* DH5α	0
S. aureus RN4220 with PRW62	*S. aureus* RN4220	171.57
S. aureus RN4220 with PRW62	PRW62	441.87
E. coli DH5α with plasmid	*E. coli* DH5α	84.96
E. coli DH5α with plasmid	*S. aureus* RN4220	0
E. coli DH5α with plasmid	PRW62	5,932.37

Table 4.5: Two-bacteria system coverage. Mean coverage per sequence for each of the two sequenced cultures

Ensuring an equal mixture of plasmid reads is possible in simple systems where plasmid copy numbers are known or controlled, but it is not easily applicable in complex metagenomic contexts. While unnormalized contig-level methylation signatures can be effective for correctly identifying a single host, it is less reliable for identifying multiple hosts.

4.3.3 Clinical Sample

Moving away from controlled samples, we wanted to examine the performance of methylation binning on a clinical stool sample, comparing the clustering results to a matched Hi-C library to provide a "ground truth". From Illumina shotgun sequencing data, we assembled metagenomic contigs. These

contigs were then binned using a matched Hi-C library, resulting in Hi-C bins, which are collections of Illumina contigs corresponding to the genetic complement of a member of the microbial community. We also sequenced the stool

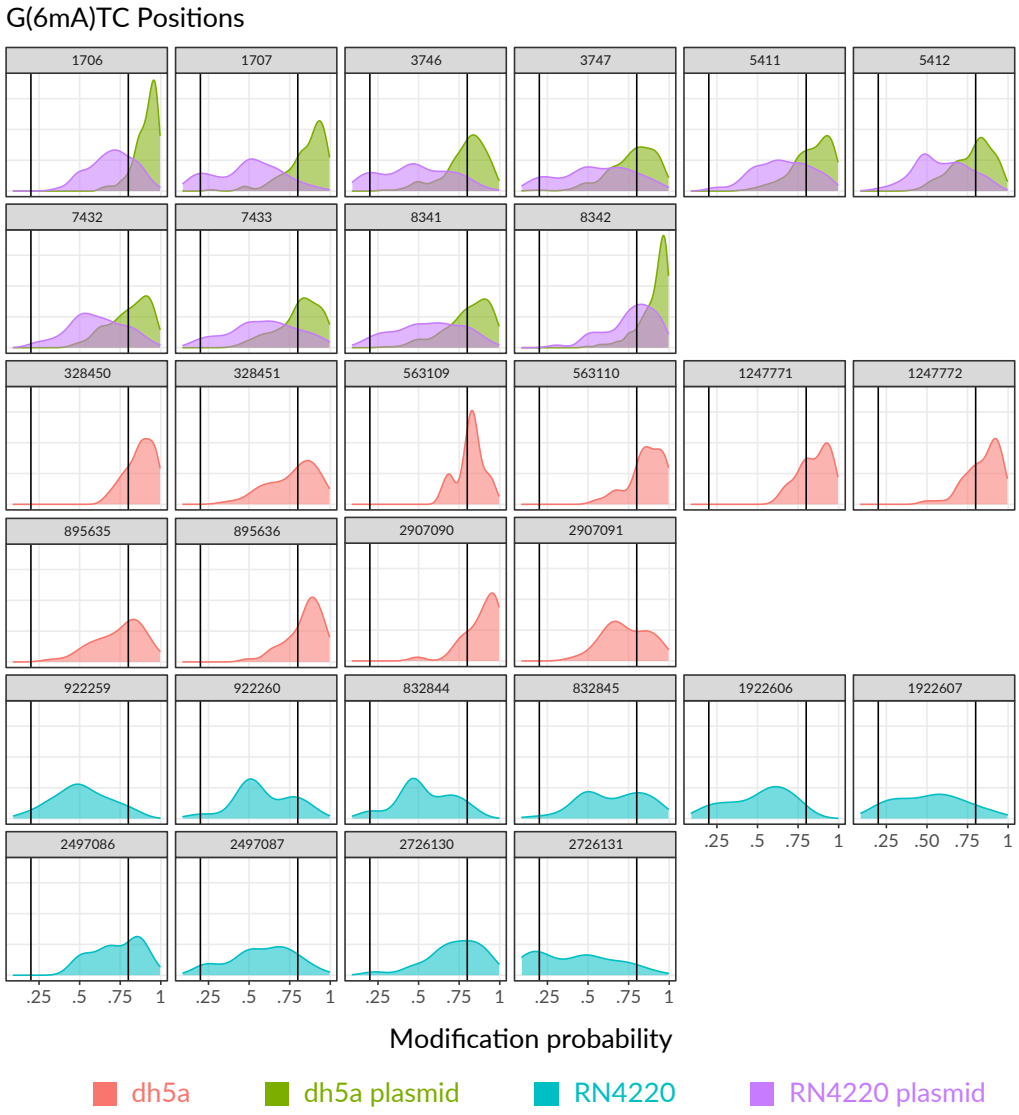

Figure 4.6: Methylation probability distributions at select dam methylation loci. Dam methylation loci, where dh5a and dh5a plasmid are methylated and RN4220 and RN4220 plasmid are not methylated. When unmethylated, the modification probability often peaks between .2 and .8, causing the majority of reads at these loci to be ignored, lowering the total number of calls for unmethylated loci (see Methods).

Chomosome	Position	Fraction of 'unclassified' methylation calls
CP017100.1 (*E. coli*)	328450	0.229
CP017100.1 (*E. coli*)	328451	0.474
CP017100.1 (*E. coli*)	563109	0.323
CP017100.1 (*E. coli*)	563110	0.2
CP017100.1 (*E. coli*)	895635	0.526
CP017100.1 (*E. coli*)	895636	0.283
CP017100.1 (*E. coli*)	1247771	0.31
CP017100.1 (*E. coli*)	1247772	0.306
CP017100.1 (*E. coli*)	2907090	0.206
CP017100.1 (*E. coli*)	2907091	0.644
PRW62 (*E. coli*)	1706	0.08
PRW62 (*E. coli*)	1707	0.252
PRW62 (*E. coli*)	3746	0.449
PRW62 (*E. coli*)	3747	0.498
PRW62 (*E. coli*)	5411	0.354
PRW62 (*E. coli*)	5412	0.469
PRW62 (*E. coli*)	7432	0.432
PRW62 (*E. coli*)	7433	0.39
PRW62 (*E. coli*)	8341	0.435
PRW62 (*E. coli*)	8342	0.134
NC_007795.1 (*S. aureus*)	832844	0.884
NC_007795.1 (*S. aureus*)	832845	0.707
NC_007795.1 (*S. aureus*)	922259	0.884
NC_007795.1 (*S. aureus*)	922260	0.826
NC_007795.1 (*S. aureus*)	1922606	0.893
NC_007795.1 (*S. aureus*)	1922607	0.861
NC_007795.1 (*S. aureus*)	2497086	0.571
NC_007795.1 (*S. aureus*)	2497087	0.833
NC_007795.1 (*S. aureus*)	2726130	0.615
NC_007795.1 (*S. aureus*)	2726131	0.667
PRW62 (*S. aureus*)	1706	0.743
PRW62 (*S. aureus*)	1707	0.814
PRW62 (*S. aureus*)	3746	0.803
PRW62 (*S. aureus*)	3747	0.767
PRW62 (*S. aureus*)	5411	0.78
PRW62 (*S. aureus*)	5412	0.861
PRW62 (*S. aureus*)	7432	0.839
PRW62 (*S. aureus*)	7433	0.783
PRW62 (*S. aureus*)	8341	0.806
PRW62 (*S. aureus*)	8342	0.514

Table 4.6: Unclassified loci. Fraction of 'unclassified' methylation calls made at select GATC loci in *S. aureus*, *E. coli* and plasmid PRW62 in each

metagenome using the ONT platform, assembled metagenomic contigs using the ONT reads, and then polished these contigs using ONT read alignments **Table** 4.7. These polished metagenomic contigs represent segments of DNA found in the microbial community, but are not binned or otherwise grouped in any way. In order to group these contigs, we called methylation and used this methylation signal to bin the long-read contigs.

Assembly	Number of contigs	N50	Longest contig	Shortest contig	Total assembly length
Polished nanopore contigs	1093	423.13 Kbp	4.61 Mbp	515 bp	128.32 Mbp
Illumina shotgun contigs	8706	52.10 Kbp	0.574 Mbp	1000 bp	116.77 Mbp

Table 4.7: Assembly summary statistics. Summary statistics for the contigs assembled with nanopore reads, and contigs assembled with Illumina reads.

Because the relevant methylation motifs in this sample are unknown, we selected the most frequently occurring ones in REBASE, as recently used by (Tourancheau et al., 2021), then narrowed these down further to the 14 motifs that occur at least once in each contig on a sufficient number of contigs, preferentially keeping 5mC motifs **Tables** 4.8**,** 4.9. We calculated a methylation signature for each contig using the percent methylation at each motif, then clustered contigs using the Euclidean distance between signatures. Polished ONT contigs were then matched to Hi-C bins using whole genome alignment.

To assess the performance of this methylation based clustering compared to Hi-C, we show the 'contamination' and 'completeness' of clusters as defined by different height thresholds on the tree, where height corresponds to methylation-based Euclidean distance. Here, 'contamination' is the percentage of contigs not contained in a cluster with only other contigs of the same

Hi-C bin, while 'completeness' is the percentage of contigs contained in the same cluster as the majority of contigs of its bin **Figure 4.7a**. Similar to a receiver operating curve (ROC) for binary classifiers, an ideal clustering would have either 0% contamination or 100% completeness at all height thresholds, resulting in an AUC of 1. Unlike an ROC, a random classifier would not be represented by a diagonal line, but a decay-like function, according to simulated random clustering.

Moving to mobile elements, we found that contigs identified as plasmids belong to the same methylation distance clusters as Hi-C bins. For all eight plasmid contigs with full barcode information, the nearest contigs come from the same bin **Figure 4.7b,c**. While these pairwise distance calculations can be done using a subset of motifs in the barcode if a plasmid contig does not

Motifs	Fraction of total contigs containing motif at least once
CAGAG	0.549
CCWGG	0.724
CMTCGAKG	0.323
CTCCAG	0.522
CTKVAG	0.821
GATC	0.659
GCCGGC	0.405
GCGC	0.749
GCWGC	0.76
GGCC	0.649
GGNNCC	0.889
GGWCC	0.745
RGCGCY	0.749
TCCGGA	0.495

Table 4.8: Clinical barcode. Motifs used in the clinical barcode, and fraction of contigs in the clinical sample containing at least one occurence each motif used for methylation binning.

contain all the motifs, the number of nearby contigs belonging to the same bin decreases as motifs are removed **Figure 4.7c**.

Using PlasmidFinder, we identified which nanopore contigs were likely to be plasmids, and using Kraken2, we classified them taxonomically. Of the nanopore contigs identified as plasmids, three were classified as *K. pneumoniae* sequences. These were an IncFIB(pQil) plasmid, an IncM2 plasmid, and a plasmid identified as both IncX3 and ColKP3. However, the IncM2

30 most commonly occuring motifs in REBASE	5mC motifs from nanodisco	Motifs from ZymoBIOMICs sample	Final Selection
GATC	GCWGC	GATC	CAGAG
CCWGG	GATC	CTKVAG	CCWGG
ATGCAT	CCGG	ANARAGTANYF	CMTCGAKG
GANTC	GATC	CTCCAG	CTCCAG
CAGAG	GGWCC	CAGAG	CTKVAG
AAGCTT	CCWGG	GCCGGC	GATC
RAATTY	GCGC	TCCGGA	GCCGGC
GGCC	GGCC	BATGCATV	GCGC
AGCCGCC	CCGCGG	CCWGG	GCWGC
CCGG	GCCGGC	CMTCGAKG	GGCC
CCTC	GGCC		GGNNCC
CATG	GGNNCC		GGWCC
CTGCAG	RGCGCY		RGCGCY
GTAC			TCCGGA
GCCGGC			
CTCGAG			
GTWWAC			
GTCGAC			
SAY			
GCGC			
CCATC			
TGGCCA			
CACAG			
GAATTC			
CAGCTG			
CTGGAG			
CCNGG			
GTNNAC			
GGTGA			
TCTAGA			

Table 4.9: Considered motifs. Methylation motifs considered for the final barcode.

and the IncX3/ColKP3 plasmids align to a Hi-C bin that is predominantly composed of *E. coli* sequences (bin_3), suggesting that *E. coli* was the host organism for these plasmids in the original sample. Methylation distance

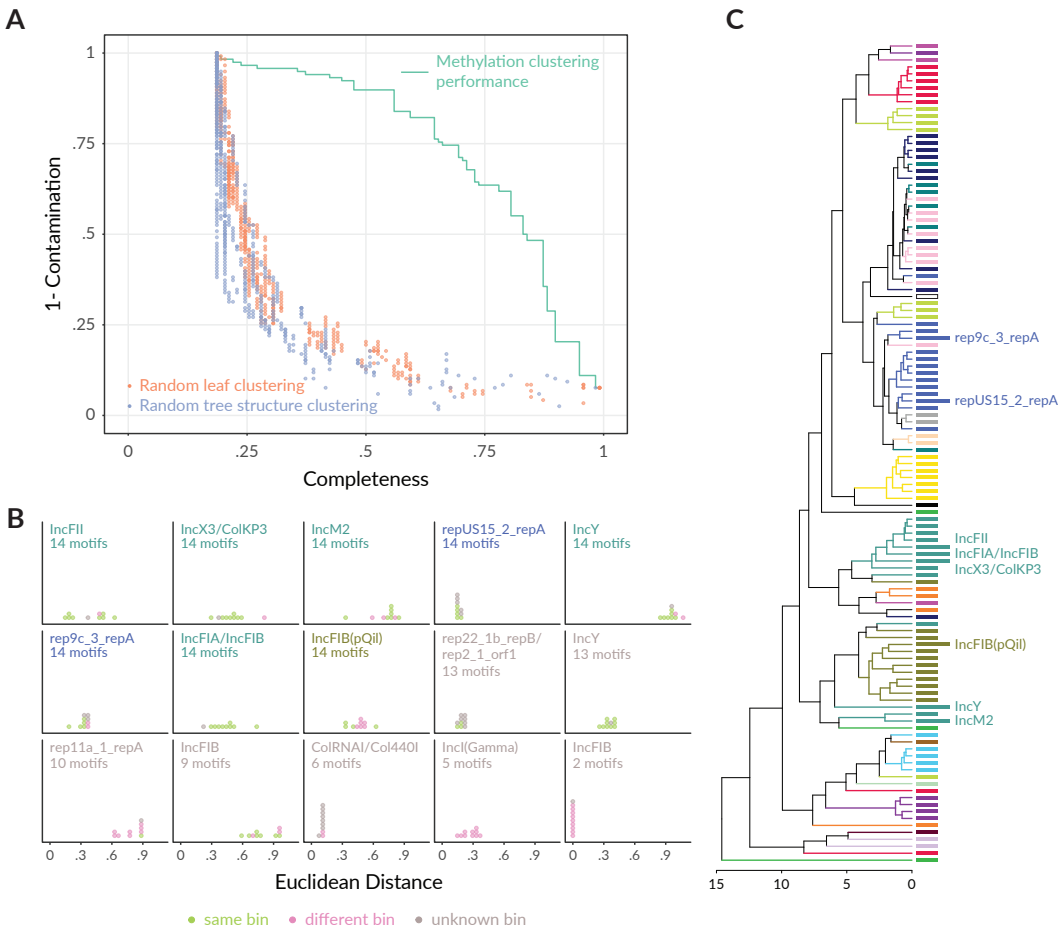

Figure 4.7: Methylation binning of a clinical sample compared to Hi-C. (A) 1-Contamination vs. Completeness curve of methylation-based hierarchical clustering of clinical metagenomic contigs. Simulations of random clustering performance are shown using points. **(B)** Tree showing hierarchical clustering of contigs, where colors represent Hi-C bins assigned to each contig. Contigs representing plasmids are shown with labels. **(C)** Distance distributions between each plasmid contig and the 20 closest contigs to each by methylation distance. Contigs are colored by whether they belong to the same Hi-C bin as the plasmid contig.

corroborates this, as the nearest contigs to these two plasmid contigs are also both classified as *E. coli* sequences. However, a ColKP3 sequence was detected in the Illumina contigs making up the *K. pneumoniae* Hi-C bin_5. While this could indicate a mis-assembly of the IncX3/ColKP3 nanopore contig, coverage profiles across each plasmid contig did not exhibit any stepwise or aberrant changes. Furthermore, the flye assembler indicates that the IncX3/ColKP3 contig is circular, suggesting that mis-assembly is unlikely, though not impossible. Without using orthogonal information from either Hi-C or methylation, correctly identifying the host of these plasmids as *E. coli* would not have been possible.

4.4 Discussion

With the drop in sequencing costs and the development of new computational tools, metagenomic sequencing has risen to be an incredibly useful tool to profile microbial samples for ecology and health research. However, problems still arise with appropriate identification of MGEs, i.e. phages, plasmids, insertion elements - where the sequence could theoretically belong to any of the hosts present in the sample. Though there are ways to profile these currently - either through culture-based or proximity ligation methods, they are costly in time and treasure. Instead, using the methylation signals embedded in nanopore sequencing data can be directly useful for metagenomic binning applications. Here we have outlined a method to do so from a typical sequencing run, without the need for any additional paired sequencing runs or data acquisition. While this work represents a starting point for these binning

applications, it is currently limited in several key ways.

Currently, non-5mCG methylation calling on the ONT platform does not consistently yield calls of the same confidence or quality as CG methylation calls. Because of this, the methylation status at many of these low confidence loci are left undetermined. We have shown that aggregation of methylation calls, both across reads and along contigs, can mitigate this issue, so that contig bin and plasmid assignments can be accurately made. This is aided by the long contigs generated from long-read metagenome assembly - longer contigs have a better chance of containing a given methylation motif. However, as methylation calling continues to improve on long-read sequencing platforms discriminatory power will also improve.

Additionally, the lack of specific methylation information presents a challenge, as constructing methylation signatures requires some prior knowledge of the motifs likely to give high differentiating power. With a highly comprehensive bacterial methylation database, motifs can be chosen based on the species and strain composition of the sample as elucidated by sequence information. While REBASE is an excellent resource to start with (Roberts et al., 2003), it lacks some 5mC motifs of common bacterial strains (Tourancheau et al., 2021). For example, the TC(5mC)GGA motif identified in the *E. coli* strain (Castellani and Chalmers 1919) of the Zymo community was not listed with any *E. coli* strain in REBASE, and the CMT(5mC)GAKG motif found in the Zymo *P. aeruginosa* (PRD-10) was not listed at all in the database. However, recent work on motif discovery has been promising (Tourancheau et al., 2021; Beaulaurier et al., 2018), and will continue to improve knowledge of

bacterial methylomes in combination with more widespread bacterial bisulfite sequencing (Oliveira Pedro H., 2021). These insights will better inform motif selection and significantly improve binning efforts, as well as further the field of bacterial epigenetics generally.

Lastly, the applicability of this method is limited because of its inability to identify plasmids in multiple hosts. However, it does not escape our notice that given the long-read nature of nanopore sequencing, it may be possible to aggregate methylation signals along whole reads, and classify them as has been done with PacBio data (Beaulaurier et al., 2018). Again, this will become more feasible as methylation calling on the platform continues to improve.

Despite these current limitations, we find that it is possible to bin some single-host plasmids on a single ONT sequencing run, which can be a useful supplement to kmer-based and coverage-based binning software. As sequencing power becomes more democratized and microbial surveillance needs become more urgent (Iskandar et al., 2021), further development of similar low-barrier metagenomic methods will be necessary.

4.5 Methods

4.5.1 Strain culture

Staphylococcus aureus RN4220 cells were grown at 37°C in Bacto Brain-Heart infusion (BHI) broth with shaking at 220 RPM. Antibiotics were used at the following concentrations for strain construction and plasmid maintenance in *S. aureus*: spectinomycin, 250 µg/mL. *Escherichia coli* Dh5a cells were grown at 37°C, unless otherwise indicated, in RPI Luria Broth (LB) with spectinomycin

at 50 µg/mL for plasmid maintenance. Strains were constructed, stored in 10% DMSO in -80°C, and restruck on BHI agar (*S. aureus*) or LB agar (*E. coli*) with spectinomycin if needed.

pRW62 was constructed by Gibson assembly using the pLZ12 backbone (primers JW713/714) and from a construct (pDB184) containing the *Streptococcus pyogenes* CRISPR-Cas locus from strain SF370 (JW715/754) with a single spacer. Plasmids were transformed into chemically competent Dh5a with heat shock or electrocompetent RN4220 with electroporation, then plated on antibiotic for selection.

4.5.2 DNA extraction and sequencing

The ZymoBIOMICS HMW DNA Standard was purchased from Zymo Research. From this, an ONT sequencing library was prepared using the ONT ligation based sequencing kit (SQK-LSK109) according to manufacturer specifications **Table** 4.4.

DNA was extracted from strains 3294 and 3689 (see Strain culture above) using the Zymo Quick-DNA Fungal Bacterial Kit, and sequenced using the ONT Rapid Sequencing Kit (SQK-RAD004) and a flongle flowcell (FLO-FLG001). For bisulfite sequencing, DNA was sheared to 300bp using the Bioruptor Pico according to manufacturer specifications. Bisulfite conversion was done using the Zymo EZ DNA Methylation Gold Kit, and a sequencing library was prepared using the Swift Accel-NGS Methyl-Seq DNA Library Kit, all according to manufacturer specifications.

DNA was extracted from a human stool sample collected at the Johns Hopkins Hospital Medical Microbiology Laboratory using the Qiagen PowerSoil DNA Isolation Kit, and an ONT library was prepared again using the ligation based sequencing kit (SQK-LSK109). These libraries were then sequenced on a FLO-MIN106D (R9.4.1) flowcell using a GridION. From the same stool sample, a proximity ligation (Hi-C) library was prepared using the Phase Genomics ProxiMeta Kit. The accompanying shotgun library was prepared also using the Qiagen PowerSoil DNA Isolation Kit for extraction, and the Illumina Nextera DNA Flex Library Prep Kit. Both the Hi-C library and the shotgun library were sequenced on a HiSeq 2500 using 2x100 chemistry.

4.5.3 Assembly and alignment

All ONT data was bascalled using the Guppy basecaller (v4.5.2) with the high accuracy model appropriate to the R9.4.1 flowcells used (dna_r9.4.1_450bps_hac). Reads with a quality score less than 9 were discarded and not used for further analysis. Any reads from the clinical stool sample aligning to the GRCh38 human reference genome using minimap2 (v2.17-r974-dirty) (Li, 2018) with the ONT data preset map-ont were also discarded and not used for further analysis. These reads were then assembled using Flye (v2.9-b1768) (Kolmogorov et al., 2020), with the —meta flag, with the genome size set to 100Mb. Using minimap2 (Li, 2018) reads were aligned back to the assembly using the ONT data preset map-ont. The assembly was then polished with the aligned reads using racon (Vaser et al., 2017)(v1.4.19) with settings -m 8 -x -6 -g -8 -w 500. To further correct the assembly, Medaka (v1.4.3) was used on default settings

with the r941_min_high_g360 model.

4.5.4 Hi-C binning and contig identification

Human reads were removed from the Hi-C data using BMTagger (v3.101) (Rotmistrovsky and Agarwala, 2011). Hi-C bins were calculated using the Phase Genomics platform ProxiWrap. Metagenomic contigs were assigned to Hi-C bins based on whole genome alignments made using Mummer (v4.0.0) (Marçais et al., 2018). For each contig, the total contig coverage by each bin was calculated, and the bin with the highest coverage was assigned to the contig. Contigs with less than 20% coverage by any bin were not assigned and excluded from bin analysis. Nanopore contigs were classified taxonomically using Kraken (v2.1.2) using the 'standard' database (Wood, Lu, and Langmead, 2019).

4.5.5 Methylation calling

Methylation was called from nanopore data with the ONT package Megalodon (v2.3.1, https://github.com/nanoporetech/megalodon), using an all-context 5mC and 6mA model (res_dna_r941_min_modbases-all-context_v001) available in Rerio (https://github.com/nanoporetech/rerio). For strains 3294 and 3689, the previously published *E. coli* NEB 5-alpha genome (accession: CP017100.1) and the *S. aureus* representative genome (accession: NC_007795.1) were used as the reference genomes for anchoring methylation calls. Genomes of the organisms included in the documentation were used as the references for the ZymoBIOMICS HMW DNA Standard. The polished metagenomic

assembly (see Assembly above) was used as the reference for the clinical stool sample.

For each read, bases with >80% methylation probability were identified as methylated, while bases with <20% methylation probability were identified as unmethylated, as per the default settings of the tool. The percent methylation at each genomic locus was determined by dividing the number of reads methylated at that locus by the total number of reads called at that locus. Reads with indeterminate methylation calls at the locus were not considered in this calculation.

4.5.6 Alignment and coverage

ONT reads were aligned to their respective reference genomes using minimap2 with preset map-ont. Coverage was calculated using the coverage module in bedtools (v2.27.1) (Quinlan and Hall, 2010).

4.5.7 Bisulfite analysis

Whole genome bisulfite sequencing (WGBS) data for each of the 7 strains included in the Zymo community standard was obtained from SRA (BioProject PRJNA477598). Bismark (v0.23.0) (Krueger and Andrews, 2011) was used to align the WGBS data for each strain to the ZymoBIOMICS reference genomes (see Methylation calling above). Genomic loci with >90% methylation frequency were considered methylated. Methylated loci exactly matching GAT5mC or C5mCWGG were counted and tabulated **Table 4.2**. Methylated loci not matching GAT5mC or C5mCWGG were visually inspected, and an

overrepresentation of TC5mCGGA and GC5mCGGC methylation context was observed in *E. coli*, while an overrepresentation of CMT(5mC)GAKG was observed in *P. aeruginosa*. These contexts were then also counted and tabulated. The remaining methylated loci were then counted and tabulated as 'unknown' methylation contexts.

Likewise in the bisulfite data of the two-bacteria system, Bismark was used to align the WGBS reads to the reference genomes, and only genomic loci with >90% methylation frequency were considered methylated. All methylated loci on the *E. coli* genome were determined to exactly match the C5mCWGG context. The two methylated loci found on the *S. aureus* genome did not match C5mCWGG and were not further contextualized.

4.5.8 Data Availability

All data is available in the European Nucleotide Archive (ENA) Study Accession PRJEB54092. All code used for analysis is freely available for download at https://github.com/timplab/fan_methbin.

References

Strous, Marc, Beate Kraft, Regina Bisdorf, and Halina E Tegetmeyer (2012). "The binning of metagenomic contigs for microbial physiology of mixed cultures". en. In: *Front. Microbiol.* 3, p. 410.

Yue, Yi, Hao Huang, Zhao Qi, Hui-Min Dou, Xin-Yi Liu, Tian-Fei Han, Yue Chen, Xiang-Jun Song, You-Hua Zhang, and Jian Tu (2020). "Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets". en. In: *BMC Bioinformatics* 21.1, p. 334.

Beaulaurier, John, Shijia Zhu, Gintaras Deikus, Ilaria Mogno, Xue-Song Zhang, Austin Davis-Richardson, Ronald Canepa, Eric W Triplett, Jeremiah J Faith, Robert Sebra, Eric E Schadt, and Gang Fang (2018). "Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation". en. In: *Nat. Biotechnol.* 36.1, pp. 61–69.

Burton, Joshua N, Ivan Liachko, Maitreya J Dunham, and Jay Shendure (2014). "Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps". en. In: *G3* 4.7, pp. 1339–1346.

Beyi, Ashenafi Feyisa, Debora Brito-Goulart, Tyler Hawbecker, Brandon Ruddell, Alan Hassall, Renee Dewell, Grant Dewell, Orhan Sahin, Qijing Zhang, and Paul J Plummer (2021). "Enrofloxacin Alters Fecal Microbiota and Resistome Irrespective of Its Dose in Calves". en. In: *Microorganisms* 9.10.

Tourancheau, Alan, Edward A Mead, Xue-Song Zhang, and Gang Fang (2021). "Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing". en. In: *Nat. Methods* 18.5, pp. 491–498.

McIntyre, Alexa B R, Noah Alexander, Aaron S Burton, Sarah Castro-Wallace, Charles Y Chiu, Kristen K John, Sarah E Stahl, Sheng Li, and Christopher E Mason (2017). "Nanopore detection of bacterial DNA base modifications". en.

Roberts, Richard J, Tamas Vincze, Janos Posfai, and Dana Macelis (2003). "REBASE: restriction enzymes and methyltransferases". en. In: *Nucleic Acids Res.* 31.1, pp. 418–420.

Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp (2017). "Detecting DNA cytosine methylation using nanopore sequencing". en. In: *Nat. Methods* 14.4, pp. 407–410.

Oliveira Pedro H. (2021). "Bacterial Epigenomics: Coming of Age". In: *mSystems* 6.4, e00747–21.

Iskandar, Katia, Laurent Molinier, Souheil Hallit, Massimo Sartelli, Timothy Craig Hardcastle, Mainul Haque, Halyna Lugova, Sameer Dhingra, Paras Sharma, Salequl Islam, Irfan Mohammed, Isa Naina Mohamed, Pierre Abi Hanna, Said El Hajj, Nurul Adilla Hayat Jamaluddin, Pascale Salameh, and Christine Roques (2021). "Surveillance of antimicrobial resistance in low- and middle-income countries: a scattered picture". en. In: *Antimicrob. Resist. Infect. Control* 10.1, p. 63.

Li, Heng (2018). "Minimap2: pairwise alignment for nucleotide sequences". en. In: *Bioinformatics* 34.18, pp. 3094–3100.

Kolmogorov, Mikhail, Derek M Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy P L Smith, and Pavel A Pevzner (2020). "metaFlye: scalable long-read metagenome assembly using repeat graphs". en. In: *Nat. Methods* 17.11, pp. 1103–1110.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić (2017). "Fast and accurate de novo genome assembly from long uncorrected reads". en. In: *Genome Res.* 27.5, pp. 737–746.

Rotmistrovsky and Agarwala (2011). "BMTagger: Best Match Tagger for removing human reads from metagenomics datasets". In: *Unpublished*.

Marçais, Guillaume, Arthur L Delcher, Adam M Phillippy, Rachel Coston, Steven L Salzberg, and Aleksey Zimin (2018). "MUMmer4: A fast and versatile genome alignment system". en. In: *PLoS Comput. Biol.* 14.1, e1005944.

Wood, Derrick E, Jennifer Lu, and Ben Langmead (2019). "Improved metagenomic analysis with Kraken 2". en. In: *Genome Biol.* 20.1, p. 257.

Quinlan, Aaron R and Ira M Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features". en. In: *Bioinformatics* 26.6, pp. 841–842.

Krueger, Felix and Simon R Andrews (2011). "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". en. In: *Bioinformatics* 27.11, pp. 1571–1572.

Chapter 5

Discussion and Conclusion

As infectious diseases continue to be of growing interest and concern (Baker et al., 2022), methods to monitor, diagnose, understand, and combat them must continue to evolve and improve. Sequencing technologies have developed at breakneck pace in recent decades (Hu et al., 2021; Schatz and Langmead, 2013), and sequencing based investigative methods have been applied to great effect in almost all fields of biology and medicine. I have taken advantage of the unique properties of the most recent generation of sequencing technology for infectious disease applications, using it to identify and surveille AMR genes, assemble eukaryotic pathogen genomes, and link plasmids to hosts in complex microbial communities.

While nanopore sequencing can detect specific AMR genes quickly and agnostically (Tamma et al., 2019), it can still be difficult to infer phenotypic resistance (Yee, Dien Bard, and Simner, 2021). Genomic data is able to provide information on an organism's potential behavior, but observations of actual activity and function require transcriptomic or proteomic data. In situ functional studies of resistance mechanisms with metatranscriptomics could help

to address these issues. As understanding of resistance mechanisms continues to grow, predictions of phenotypic resistance will become increasingly accurate and clinically actionable.

Use of long read sequencing for genome assembly has unlocked continuously larger genomes (Neale et al., 2014), and accompanying software has made high quality genome assembly more accessible than ever before (Fan et al., 2021). As more and more eukaryotic pathogen genomes are sequenced, collated, and curated (Aurrecoechea et al., 2017), they will become crucial to the development of sequencing-based diagnostics of infectious diseases (Lu and Salzberg, 2018), in addition to being useful for furthering basic science research on these organisms.

Long-read data for metagenomic assembly not only produces longer contigs, but is able to preserve base modification data, which can be used for binning applications. Not only can contigs be grouped on the bases of base modifications, but reads can as well. Although this has not been shown using a single nanopore run, its possibility has been demonstrated using PacBio data (Beaulaurier et al., 2018). Developing this capability in nanopore sequencing would enable multi-host plasmid assignments which are currently unfeasible.

References

Baker, Rachel E, Ayesha S Mahmud, Ian F Miller, Malavika Rajeev, Fidisoa Rasambainarivo, Benjamin L Rice, Saki Takahashi, Andrew J Tatem, Caroline E Wagner, Lin-Fa Wang, Amy Wesolowski, and C Jessica E Metcalf (2022). "Infectious disease in an era of global change". en. In: *Nat. Rev. Microbiol.* 20.4, pp. 193–205.

Hu, Taishan, Nilesh Chitnis, Dimitri Monos, and Anh Dinh (2021). "Next-generation sequencing technologies: An overview". en. In: *Hum. Immunol.* 82.11, pp. 801–811.

Schatz, Michael C and Ben Langmead (2013). "The DNA Data Deluge: Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze". en. In: *IEEE Spectrum* 50.7, pp. 26–33.

Tamma, P D, Y Fan, Y Bergman, G Pertea, and others (2019). "Applying rapid whole-genome sequencing to predict phenotypic antimicrobial susceptibility testing results among carbapenem-resistant *Klebsiella pneumoniae* clinical isolates". In: *Antimicrob. Agents Chemother.*

Yee, Rebecca, Jennifer Dien Bard, and Patricia J Simner (2021). "The Genotype-to-Phenotype Dilemma: How Should Laboratories Approach Discordant Susceptibility Results?" en. In: *J. Clin. Microbiol.* 59.6.

Neale, David B, Jill L Wegrzyn, Kristian A Stevens, Aleksey V Zimin, Daniela Puiu, Marc W Crepeau, Charis Cardeno, Maxim Koriabine, Ann E Holtz-Morris, John D Liechty, Pedro J Martínez-García, Hans A Vasquez-Gross, Brian Y Lin, Jacob J Zieve, William M Dougherty, Sara Fuentes-Soriano, Le-Shin Wu, Don Gilbert, Guillaume Marçais, Michael Roberts, Carson Holt, Mark Yandell, John M Davis, Katherine E Smith, Jeffrey F D Dean, W Walter Lorenz, Ross W Whetten, Ronald Sederoff, Nicholas Wheeler, Patrick E McGuire, Doreen Main, Carol A Loopstra, Keithanne Mockaitis, Pieter J deJong, James A Yorke, Steven L Salzberg, and Charles H Langley (2014). "Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies". en. In: *Genome Biol.* 15.3, R59.

Fan, Yunfan, Andrew N Gale, Anna Bailey, Kali Barnes, Kiersten Colotti, Michal Mass, Luke B Morina, Bailey Robertson, Remy Schwab, Niki Tselepidakis, and Winston Timp (2021). "Genome and transcriptome of a pathogenic yeast, *Candida nivariensis*". en. In: *G3 Genes | Genomes | Genetics* 11.7.

Aurrecoechea, Cristina, Ana Barreto, Evelina Y Basenko, John Brestelli, Brian P Brunk, Shon Cade, Kathryn Crouch, Ryan Doherty, Dave Falke, Steve Fischer, Bindu Gajria, Omar S Harb, Mark Heiges, Christiane Hertz-Fowler, Sufen Hu, John Iodice, Jessica C Kissinger, Cris Lawrence, Wei Li, Deborah F Pinney, Jane A Pulman, David S Roos, Achchuthan Shanmugasundram, Fatima Silva-Franco, Sascha Steinbiss, Christian J Stoeckert Jr, Drew Spruill, Haiming Wang, Susanne Warrenfeltz, and Jie Zheng (2017). "EuPathDB: the eukaryotic pathogen genomics database resource". en. In: *Nucleic Acids Res.* 45.D1, pp. D581–D591.

Lu, Jennifer and Steven L Salzberg (2018). "Removing contaminants from databases of draft genomes". en. In: *PLoS Comput. Biol.* 14.6, e1006277.

Beaulaurier, John, Shijia Zhu, Gintaras Deikus, Ilaria Mogno, Xue-Song Zhang, Austin Davis-Richardson, Ronald Canepa, Eric W Triplett, Jeremiah J Faith, Robert Sebra, Eric E Schadt, and Gang Fang (2018). "Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation". en. In: *Nat. Biotechnol.* 36.1, pp. 61–69.

3400 N. Charles St.
Clark Hall 107
Baltimore, MD 21218

Yunfan Fan

(she/her)

yfan2012@gmail.com
github: yfan2012
(440)525-3050

TECHNICAL SKILLS

Computational skills: Highly proficient with R, python, bash scripting, AWS, slurm, SGE, version control (git). Very experienced using sequencing analysis software tools for read alignment, genome assembly, variant calling, SV calling, transcriptome assembly/quantification, methylation analysis, metagenomics, public data access, etc. Competent with MATLAB, LaTeX, Adobe Illustrator.
Lab techniques: Expertise in DNA/RNA extraction and handling, NGS-based assays, nanopore library prep and sequencing (ONT), mammalian cell culture.

EDUCATION

Johns Hopkins University Baltimore, MD
Ph.D. Biomedical Engineering September 2022
Advisor - Dr. Winston Timp
Thesis: Nanopore sequencing for infectious disease applications

Johns Hopkins University Baltimore, MD
B.S. Biomedical Engineering (major), French Literature (minor) May 2016
B.A. Electrical Engineering (major)

AWARDS AND HONORS

Doctoral Foreign Study Award 2018 – 2021
Three-year graduate fellowship awarded by the Canadian Institutes of Health Research (CIHR) providing special recognition and support to Canadian students pursuing a doctoral degree in a health-related field outside of Canada.
Title: Clinical infectious disease sequencing for antimicrobial resistance detection and antibiotic stewardship

Provost's Undergraduate Research Award (PURA) 2015
Supports and encourages Hopkins undergraduate students to engage in independent research, scholarly and creative projects.
Title: Methylation Sequencing on the MinION

Irini J. Maroulis Award 2015
Awarded each year to a female undergraduate student in the Johns Hopkins University Whiting School of Engineering for outstanding community service and outreach.

PUBLICATIONS

1. **Y. Fan**, A. N. Gale, A. Bailey, K. Barnes, K. Colotti, M. Mass, L. B. Morina, B. Robertson, R. Schwab, N. Tselepidakis, and W. Timp, "Genome and transcriptome of a pathogenic yeast, *Candida nivariensis*," *G3: Genes, Genomes, Genetics*, vol. 11, no. 7, p. jkab137, 2021

2. J. Vornhagen, C. M. Bassis, S. Ramakrishnan, R. Hein, S. Mason, Y. Bergman, N. Sunshine, **Y. Fan**, C. L. Holmes, W. Timp, M. C Schatz, V. B Young, P. J. Simner, and M. A. Bachman, "A plasmid locus associated with *Klebsiella* clinical infections encodes a microbiome-dependent gut fitness factor," *PLoS Pathogens*, vol. 17, no. 4, p. e1009537, 2021

3. S. Kovaka, **Y. Fan**, B. Ni, W. Timp, and M. C. Schatz, "Targeted nanopore sequencing by real-time mapping of raw electrical signal with uncalled," *Nature Biotechnology*, vol. 39, no. 4, pp. 431–441, 2021

4. P. M. Thielen, S. Wohl, T. Mehoke, S. Ramakrishnan, M. Kirsche, O. Falade-Nwulia, N. S. Trovão, A. Ernlund, C. Howser, N. Sadowski, C. P. Morris, M. Hopkins, M. Schwartz, **Y. Fan**, V. Gniazdowski, J. Lessler, L. Sauer, M. C. Schatz, J. D. Evans, S. C. Ray, W. Timp, and H. H. Mostafa, "Genomic diversity of sars-cov-2 during early introduction into the baltimore–washington metropolitan area," *JCI Insight*, vol. 6, no. 6, 2021

5. A. Gershman, T. G. Romer, **Y. Fan**, R. Razaghi, W. A. Smith, and W. Timp, "De novo genome assembly of the tobacco hornworm moth (*Manduca sexta*)," *G3: Genes, Genomes, Genetics*, vol. 11, no. 1, p. jkaa047, 2021

6. J. L. Drewes, A. Corona, U. Sanchez, **Y. Fan**, S. K. Hourigan, M. Weidner, S. D. Sidhu, P. J. Simner, H. Wang, W. Timp, M. Oliva-Hemker, and C. L. Sears, "Transmission and clearance of potential procarcinogenic bacteria during fecal microbiota transplantation for recurrent *Clostridioides difficile*," *JCI Insight*, vol. 4, no. 19, 2019

7. P. D. Tamma, **Y. Fan**, Y. Bergman, G. Pertea, A. Q. Kazmi, S. Lewis, K. C. Carroll, M. C. Schatz, W. Timp, and P. J. Simner, "Applying rapid whole-genome sequencing to predict phenotypic antimicrobial susceptibility testing results among carbapenem-resistant *Klebsiella pneumoniae* clinical isolates," *Antimicrobial Agents and Chemotherapy*, vol. 63, no. 1, pp. e01923–18, 2019

8. V. Beleva Guthrie, D. L. Masica, A. Fraser, J. Federico, **Y. Fan**, M. Camps, and R. Karchin, "Network analysis of protein adaptation: modeling the functional impact of multiple mutations," *Molecular Biology and Evolution*, vol. 35, no. 6, pp. 1507–1519, 2018

9. P. D. Tamma, **Y. Fan**, Y. Bergman, A. C. Sick-Samuels, A. J. Hsu, W. Timp, P. J. Simner, B. C. Prokesch, and D. E. Greenberg, "Successful treatment of per-

sistent *Burkholderia cepacia* complex bacteremia with ceftazidime-avibactam," *Antimicrobial Agents and Chemotherapy*, vol. 62, no. 4, pp. e02213–17, 2018

10. R. Luo, A. Zimin, R. Workman, **Y. Fan**, G. Pertea, N. Grossman, M. P. Wear, B. Jia, H. Miller, A. Casadevall, W. Timp, S. X. Zhang, and S. L. Salzberg, "First draft genome sequence of the pathogenic fungus *Lomentospora prolificans* (formerly *Scedosporium prolificans*)," *G3: Genes, Genomes, Genetics*, vol. 7, no. 11, pp. 3831–3836, 2017

11. J. J. Credle, C. Y. Itoh, T. Yuan, R. Sharma, E. R. Scott, R. E. Workman, **Y. Fan**, F. Housseau, N. J. Llosa, W. R. Bell, H. Miller, S. X. Zhang, W. Timp, and H. B. Larman, "Multiplexed analysis of fixed tissue rna using ligation in situ hybridization," *Nucleic Acids Research*, vol. 45, no. 14, pp. e128–e128, 2017

12. A. L. Norris, R. E. Workman, **Y. Fan**, J. R. Eshleman, and W. Timp, "Nanopore sequencing detects structural variants in cancer," *Cancer Biology & Therapy*, vol. 17, no. 3, pp. 246–253, 2016

PRESENTATIONS

Conference Talks

SFAF - Sequencing, Finishing and Analysis in the Future (2018) Santa Fe, NM
Title: Bacterial sequencing and assembly for analysis of antibiotic resistance genes and mutations

Nanopore Community Meeting (2017) New York, NY
Title: Bacterial DNA modifications

Poster Presentations

AGBT - Advances in Genome Biology and Technology (2020) Marco Island, FL
Title: Calling bacterial methylation signatures using nanopore sequencing

Genome Informatics (2019) Cold Spring Harbor Laboratory
Title: Genome assembly using R9 and R10 type ONT flowcells

ASM Next-Generation Sequencing (2018) Washington, D.C.
Title: Nanopore sequencing for AMR detection and characterization