ABSTRACT

Title of Dissertation: ALGORITHMS FOR SCALABLE AND EFFICIENT POPULATION GENOMICS AND METAGENOMICS

Kiran Gajanan Javkar
Doctor of Philosophy, 2022

Dissertation Directed by: Professor Mihai Pop
Department of Computer Science

Microbes strongly impact human health and the ecosystem of which they are a part. Rapid improvements and decreasing costs in sequencing technologies have revolutionized the field of genomics and enabled important insights into microbial genome biology and microbiomes. However, new tools and approaches are needed to facilitate the efficient analysis of large sets of genomes and to associate genomic features with phenotypic characteristics better. Here, we built and utilized several tools for large-scale whole-genome analysis for different microbial characteristics, such as antimicrobial resistance and pathogenicity, that are important for human health.

Chapters 2 and 3 demonstrate the needs and challenges of population genomics in associating antimicrobial resistance with genomic features. Our results highlight important limitations of reference database-driven analysis for genotype-phenotype association studies and demonstrate the utility of whole-genome population genomics in uncovering novel genomic factors associated with antimicrobial

resistance.

Chapter 4 describes PRAWNS, a fast and scalable bioinformatics tool that generates compact pan-genomic features. Existing approaches are unable to meet the needs of large-scale whole-genome analyses, either due to scalability limitations or the inability of the genomic features generated to support a thorough whole-genome assessment. We demonstrate that PRAWNS scales to thousands of genomes and provides a concise collection of genomic features which support the downstream analyses.

In Chapter 5, we assess whether the combination of long and short-read sequencing can expedite the accurate reconstruction of a pathogen genome from a microbial community. We describe the challenges for pathogen detection in current foodborne illness outbreak monitoring. Our results show that the recovery of a pathogen genome can be accelerated using a combination of long and short-read sequencing after limited culturing of the microbial community. We evaluated several popular genome assembly approaches and identified areas for improvement.

In Chapter 6, we describe SIMILE, a fast and scalable bioinformatics tool that enables the detection of genomic regions shared between several assembled metagenomes. In metagenomics, microbial communities are sequenced directly without culturing. Although metagenomics has furthered our understanding of the microbiome, comparing metagenomic samples is extremely difficult. We describe the need and challenges in comparing several metagenomic samples and present an approach that facilitates large-scale metagenomic comparisons.

ALGORITHMS FOR SCALABLE AND EFFICIENT POPULATION GENOMICS AND METAGENOMICS

by

Kiran Gajanan Javkar

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Mihai Pop, Chair
Dr. Hugh Rand
Professor Rob Patro
Professor John Dickerson
Professor Brantley Hall (Dean's Representative)

Preface

The bioinformatic analyses, algorithms, tools, and results have either been published in peer-reviewed journals or are currently under submission. At the time of this writing, Chapters 2, 3, and 5 have been published, Chapter 4 has been submitted and is currently under review, and Chapter 5 is currently under preparation for submission. Each chapter has been reformatted here. I am thankful for the help of all the co-authors; their knowledge, expertise, and dedication greatly improved my understanding of the projects and the overall quality of my research.

- **Chapter 2**
    - **Kiran Javkar**, Hugh Rand, Maria Hoffmann, Yan Luo, Saul Sarria, Nagarajan Thirunavukkarasu, Christine A Pillai, Patrick McGann, J Kristie Johnson, Errol Strain, and Mihai Pop. "Whole-genome assessment of clinical *Acinetobacter baumannii* isolates uncovers potentially novel factors influencing carbapenem resistance." *Frontiers in microbiology* 12 (2021). My contributions include (1) literature review, (2) experimental design, (3) running experiments, (4) data analysis, and (5) writing the manuscript.
- **Chapter 3**

- Jay Noboru Worley, **Kiran Javkar**, Maria Hoffmann, Kristen Hysell, Amanda Garcia-Williams, Kaitlin Tagg, Sanjat Kanjilal, Errol Strain, Mihai Pop, Marc Allard, Louise Francois Watkins, and Lynn Bry. "Genomic Drivers of Multidrug-Resistant *Shigella* Affecting Vulnerable Patient Populations in the United States and Abroad." *Mbio* 12.1 (2021): e03188-20. My contributions include (1) experimental design, (2) running experiments, (3) data analysis, and (4) writing the manuscript.

- **Chapter 4**

  - **Kiran Javkar**, Hugh Rand, Errol Strain, and Mihai Pop. "PRAWNS: Compact pan-genomic features for whole-genome population genomics." *Manuscript under review* (2022). My contributions include (1) literature review, (2) experimental design, (3) design and implementation of the algorithms, (4) running experiments, (5) data analysis, (6) updating and maintaining the software and (7) writing the manuscript.

- **Chapter 5**

  - Seth Commichaux*, **Kiran Javkar***, Padmini Ramachandran, Niranjan Nagarajan, Denis Bertrand, Yi Chen, Elizabeth Reed, Narjol Gonzalez-Escalona, Errol Strain, Hugh Rand, Mihai Pop, and Andrea Ottesen. "Evaluating the accuracy of *Listeria monocytogenes* assemblies from quasi-metagenomic samples using long and short reads." *BMC genomics* 22.1 (2021): 1-18. My contributions include (1) literature review, (2) ex-

perimental design, (3) running experiments, (4) data analysis, and (5) writing the manuscript.

- **Chapter 6**
  - **Kiran Javkar**, Hirak Sarkar, Hugh Rand, Rob Patro, and Mihai Pop. "SIMILE: Discover similar genomic regions shared across a collection of metagenomic samples." *Manuscript under preparation for submission.* My contributions include (1) literature review, (2) experimental design, (3) design and implementation of the algorithms, (4) running experiments, (5) data analysis, (6) updating and maintaining the software and (7) writing the manuscript.

Following are other publications to which I have contributed during my PhD:

- Jacquelyn S Meisel, Daniel J Nasko, Brian Brubach, Victoria Cepeda-Espinoza, Jessica Chopyk, Héctor Corrada-Bravo, Marcus Fedarko, Jay Ghurye, **Kiran Javkar**, Nathan D Olson *et al.* "Current progress and future opportunities in applications of bioinformatics for biodefense and pathogen detection: report from the Winter Mid-Atlantic Microbiome Meet-up, College Park, MD, January 10, 2018." *Microbiome* 6.1 (2018): 1-10.

- Suraj Nair*, **Kiran Javkar***, Jiahui Wu, and Vanessa Frias-Martinez. "Understanding cycling trip purpose and route choice using GPS traces and open data." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.1 (2019): 1-26.

- Debjani Saha, Anna Chan, Brook Stacy, **Kiran Javkar**, Sushant Patkar, and Michelle L Mazurek. "User attitudes on direct-to-consumer genetic testing." *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE (2020).

- Joan Martí-Carreras, Alejandro Rafael Gener, Sierra D Miller, Anderson F Brito, Christiam E Camacho, Ryan Connor *et al.* "NCBI's virus discovery codeathon: building "FIVE"—The Federated Index of Viral Experiments API index." *Viruses* 12.12 (2020): 1424

- Seth Commichaux, **Kiran Javkar**, Harihara Muralidharan, Padmini Ramachandran, Andrea Ottesen, and Mihai Pop. "taxaTarget: Fast, sensitive, and precise classification of microeukaryotes in metagenomic data." *Preprint on Research Square* (2022).

- Seth Commichaux, Hugh Rand, **Kiran Javkar**, Erin K. Molloy, James B. Pettengill, Arthur Pightling, Maria Hoffmann, Mihai Pop, Victor Jayeola, Steven Foley, and Yan Luo. "Assessing the use of plasmids for relating the 2020 *Salmonella enterica* serovar Newport onion outbreak to the farms implicated by the outbreak investigation." *Manuscript under preparation for submission* (2022).

Dedication

To my parents and my beloved partner

Acknowledgments

First and foremost, I want to thank my advisor, Prof. Mihai Pop, and my mentors Dr. Hugh Rand and Dr. Errol Strain for helping me immensely throughout my PhD. Despite having no prior research background, Mihai allowed me to join his lab and introduced me to Hugh and Errol. I have had interesting scientific conversations with all three of them, which helped me improve my understanding of biology and also gave me a sense of intellectual freedom and perspective to pursue research. I am filled with immense gratitude for their never-ending support, motivation, patience, and guidance over all these years.

I am extremely grateful to my committee members Prof. Rob Patro, Prof. John Dickerson, and Prof. Brantley Hall for their continued support of my research. Our discussions and their feedback have been invaluable in exploring new research ideas and shaping the overall dissertation.

I have been fortunate to have networked, brain-stormed, and collaborated with numerous amazing researchers. These include the current, past, and honorary members of the Pop lab, CBCB, and the FDA CFSAN—Domenick Braccia, Brian Brubach, Victoria Cepeda-Espinoza, Seth Commichaux, Héctor Corrada-Bravo, Steve Davis, Maria Hoffmann, Tu Luan, Yan Luo, Shiva Mehravaran, Jacquelyn S Michaelis, Harihara Muralidharan, Dan Nasko, Nate Olson, James Pettengill,

Arthur Pightling, Hirak Sarkar, Jeremy Selengut, Dylan Taylor, Todd Treangen, and Mohsen Zakeri. I have relied on many of these colleagues for endless talks, venting, and even a good laugh. I am grateful to have the opportunity to work with some phenomenal collaborators from other institutes as well—Patrick McGann, Jay N Worley, and Lynn Bry. I would like to thank Prof. Vanessa Friaz-Martinez who provided me with my first ever research experience at UMD. I am eternally grateful to Prof. Rajiv Gandhi, without whom I would not have envisioned pursuing a PhD.

This journey would not have been possible without the support and cooperation of the CBCB, Department of Computer Science, UMIACS, JIFSAN, and FDA CFSAN—including the fellow students, faculty, and staff. Special thanks to Tom Hurst, Barbara Lewis, Erica Mudd, Mary Grimley, and Prof. Jianghong Meng who ensured that I could focus on my research without worrying about external hassles. I would also like to thank Dr. Susan Martin whose guidance helped me enormously with the transitions post-PhD.

Finally, I would like to thank my family and friends for all their love and support. And my partner, Shivangi Borate, I cannot thank her enough for being my rock-solid support through all the highs and lows of this journey; I could not make it across the finish line without her unconditional love.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AMR	**A**nti**m**icrobial **R**esistance
AST	**A**ntimicrobial **S**usceptibility **T**esting
bp	**b**ase **p**air
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
CARD	**C**omprehensive **A**ntibiotic **R**esistance **D**atabase
CLSI	**C**linical & **L**aboratory **S**tandards **I**nstitute
CRAB	**C**arbapenem **R**esistant ***A****cinetobacter* ***b****aumannii*
dBg	**d**e **B**ruijn **g**raph
FDA	**F**ood and **D**rug **A**dministration
GWAS	**G**enome **W**ide **A**ssociation **S**tudies
HGT	**H**orizontal **G**ene **T**ransfer
Indel	**In**sertion (and) **D**eletion
IS	**I**nsertion **S**equence
LCB	**L**ocally **C**ollinear **B**lock
MAG	**M**etagenome **A**ssembled **G**enome
MDR	**M**ulti **D**rug **R**esistant
MGE	**M**obile **G**enetic **E**lement
MIC	**M**inimum **I**nhibitory **C**oncentration
MLST	**M**ulti**l**ocus **S**equence **T**yping
MSM	**M**en who have **S**ex with **M**en
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
NGS	**N**ext **G**eneration **S**equencing
nt	**n**ucleo**t**ide
SNP	**S**ingle **N**ucleotide **P**olymorphism
WGS	**W**hole **G**enome **S**equencing

Chapter 1: Introduction

Microbes have compelling impacts on human and ecological health [2]. For instance, microbes are present across different habitats and can influence the environmental diversity [3], some of these microbes have salient contributions to industrial processes (e.g. in apple cider vinegar and fermented apple beverages production [4]), and some microbes are commensal or vital to humans [5]. In contrast, other microbes are pathogenic—result in infectious diseases—and can resist the actions of different drugs [6, 7]. These diverse characteristics exhibited by different microbes can be better understood through the analysis of their genomes. In this thesis, we explore the development and applications of various approaches for microbial genome analysis and identify the genomic factors relevant to human health.

Microbial genome analysis starts with the process of genome sequencing. Figure 1.1 displays the overview of stages in genome sequencing and analysis of a bacterial isolate. First, the DNA is extracted out of the cell and is subjected to library preparation steps to get shredded DNA. Next, the shredded DNA library is provided to a sequencing machine that generates sequencing reads (DNA reads). These reads are assembled to construct an 'assembled genome', which is often a col-

lection of fragmented contiguous DNA sequences called contigs [8]. The reads and assembled genomes form the basis for many downstream bioinformatics analyses.

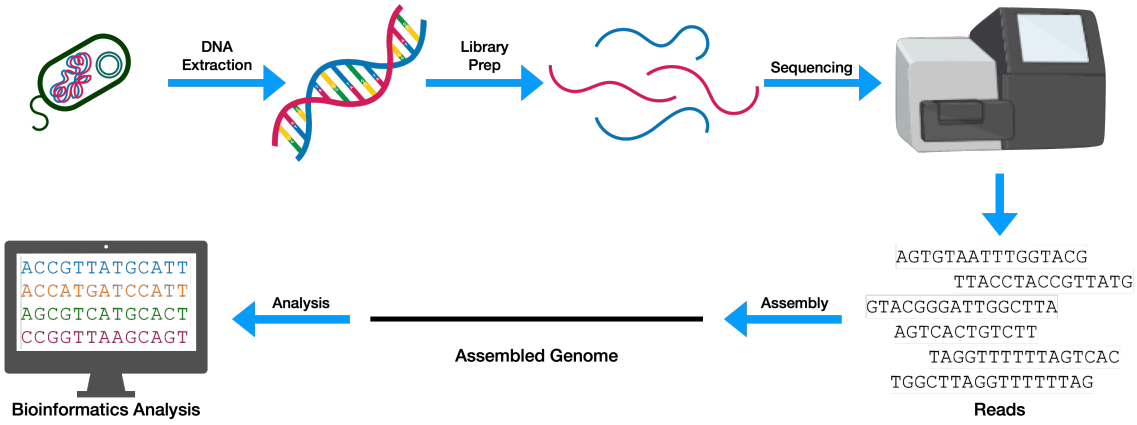


Figure 1.1: Overview of bacterial genome sequencing and analysis.

The first complete genome sequenced from free-living bacteria was obtained in 1995 [9]. It sparked an interest to sequence other bacterial species and strains that were closely related to those that were already sequenced and compare the corresponding genomes to uncover genome-wide similarities and differences. In recent times, the emergence and rapid improvements in next-generation sequencing (NGS) technologies have revolutionized the field and have resulted in the increasing availability of sequenced organisms worldwide. The availability of large number of whole-genome sequenced isolates has given rise to the field of population genomics and comparative genomics.

1.1 Population genomics

Population genomics broadly refers to the simultaneous study of numerous genomic regions from several genomes to better understand the evolutionary processes that influence the variations across genomes and populations [10]. Genomic comparisons of related organisms can provide valuable insights into their genome evolution and allow the characterization of genomic variations, such as horizontal gene transfer (HGT) events (i.e. movement of genetic material between contemporary organisms rather than 'vertical' transmission of genetic material from parent/s to offspring via reproduction), insertions-deletions (indels), and translocations. Such variations can affect bacterial pathogenicity or virulence and the organismal survivability in different environments [1, 11]. Population genomics facilitates large-scale microbial genome analyses to gain insights into such genomic variations [12]; it is transforming public health and food safety monitoring and has led to the development of sequence-based infectious disease programs [13, 14].

Pan-genome representation, i.e., a single unified representation that aggregates the genomic features present across several closely related genomes [15], has been instrumental in evaluating the influence of genomic features on the observable differences between related organisms. For instance, some isolates of a bacterial species could be resistant to an antimicrobial drug, while others of the same species could be susceptible to that drug [16, 17]. Similarly, some bacterial isolates could be pathogenic to humans, while other related ones are non-pathogenic

or even beneficial [18, 19]. Such observable characteristics exhibited by the isolates are referred to as the 'phenotypes'; the genomic factors influential or associated with these characteristics—such as genes acquired by horizontal gene transfer, genomic rearrangements, mobile elements, other genomic mutations—are referred to as the 'genotypes'. Identification of such associations between the genomic factors and the observed characteristics constitutes a genotype-phenotype association study. Such studies permit the association of different phenotypes to their corresponding genomic variations [20, 21, 22, 23]). Identifying such genomic variations and understanding the mechanisms underlying the phenotypic variations have been imperative in monitoring the phenotypes in public health applications [24].

1.2 Antimicrobial resistance (AMR)

One of the crucial applications of genotype-phenotype correlation studies is the detection and tracking of antimicrobial resistance (AMR). Antimicrobial drugs are small molecules that inhibit the growth or kill bacteria, and have been commonly used as therapeutics for several bacterial infections [1]. Some bacteria can, however, survive and grow despite the antimicrobial pressures—this property is referred to as antimicrobial resistance. The rate of antimicrobial resistance among pathogens has been increasing worldwide and has a significant impact on public health—an estimated 700,000 deaths annually worldwide [25, 26]. A sizable proportion of AMR bacterial infections are hospital-acquired and pose a high healthcare

burden in developed and developing countries alike [27]. Some of these bacterial infections are resistant to even the most powerful antibiotics available; including the ones typically considered to be the 'last-line' antibiotics used in the treatment of critically-ill patients [28]. A bacterial isolate can be labelled as multi-drug resistant (MDR), extensively-drug-resistant (XDR), or pan-drug-resistant (PDR) based on the number of antibiotics classes the organism is resistant to.

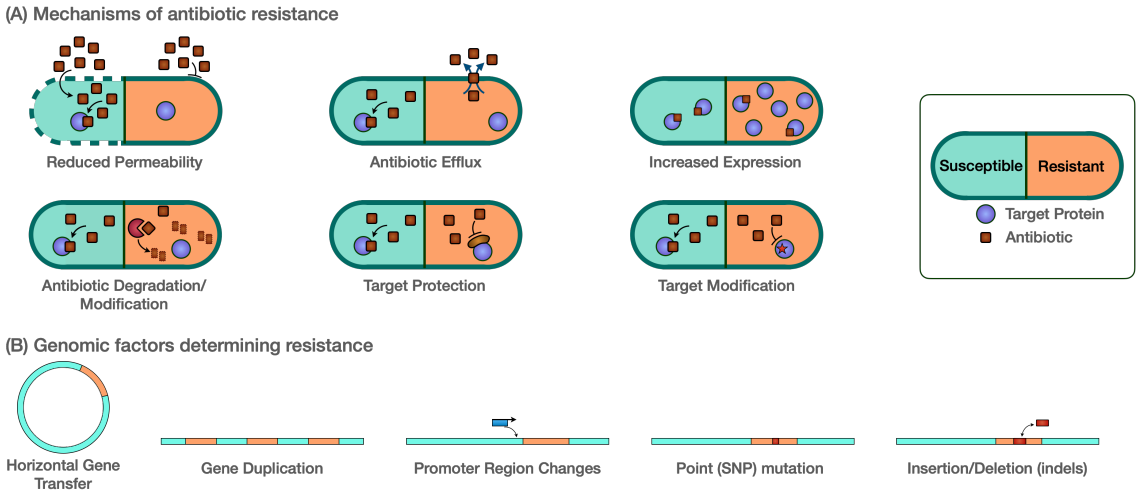


Figure 1.2: Antimicrobial resistance mechanisms and genomic factors. (A) Mechanisms of antimicrobial resistance with bacteria—susceptible organisms are represented on the left while the resistant ones depicted on the right. (B)| Genomic factors driving the antimicrobial resistance. [1]

Antimicrobial resistance is conferred by a wide variety of genetic mechanisms that vary between and within species. To develop new effective antimicrobial agents (antibiotics), we must understand how existing antimicrobials work and the bacterial mechanisms of resistance. Antimicrobial agents 'target' vital microbial functions. Different antibiotics undertake different actions to kill/inhibit the bacterial growth and can be broadly categorized as follows [29]: (i) Interfering with bacterial cell wall synthesis—cell wall enables bacteria to maintain cell shape and protection

from lysis due to intracellular osmotic pressure (ii) Inhibiting protein synthesis (i.e., blocking transcription and/or translation) (iii) Inhibiting nucleic acid synthesis to inhibit DNA replication or resulting in abnormal DNA formation (iv) Inhibition of metabolic pathways/bacterial enzymes (v) Interrupting the bacterial cell membrane structure, which may cause in osmotic imbalance resulting in cell death. Resistance has been observed to nearly all types of antibiotics via a variety of AMR mechanisms; Figure 1.2 (part (A)) illustrates some of these AMR mechanisms: (i) Reduced permeability is when the antibiotic is unable to enter the bacteria and, hence, cannot prevent bacterial systhesis via abovementioned actions. (ii) With antibiotic efflux, isolates upregulate their efflux pumps and discard the antibiotics. (iii) Isolates can increase the expression levels of different genes; if the target protein is highly expressed, then the administered dosage of antibiotics is insufficient to bind to drug targets and inhibit the bacterial growth. (iv) Isolates can encode other genes which can disintegrate or degrade the antibiotics and prevent their inhibitor actions. (v) Other genes could be encoded that enable target site protection, shielding them from the antibiotics. (vi) The target site, itself, undergoes some modifications and antibiotics are unable to bind to new target site. Each AMR mechanism has its genomic underpinnings. Figure 1.2 (part (B)) depicts the genomic factors that encode and drive these AMR mechanisms. The discovery and understanding of these genomic factors have been accelerated by advancements in whole genome sequencing technologies and development of methods for analyzing the resulting data [1].

The genome of a bacterial isolate can encode several AMR mechanisms which

can confer resistance to multiple antimicrobial drugs [30, 31]. The genomic underpinnings of these AMR mechanisms can be present on the bacterial chromosome or extra-chromosomal DNA molecules, like the plasmids. Plasmids facilitate horizontal gene transfer between bacteria and often encode genes, like the AMR factors, that help bacteria survive local environmental stress. Even a susceptible isolate can host several AMR mechanisms, suggesting the presence of AMR associated genomic factors is not always sufficient to confer antibiotic resistance. Similarly, the absence of certain strong AMR factors does not necessarily imply the corresponding bacterial isolate to exhibit a susceptible phenotype. It is extremely challenging to accurately identify the genomic factors and their associated resistance mechanisms that are causal for the antimicrobial resistance phenotype. An accurate determination of the encoded AMR mechanisms can assist in creating treatment therapies for the bacterial infections and monitoring their spread.

Population genomics has shown enormous promise to expedite the discovery of known and novel genomic factors associated with antimicrobial resistance from large genomic collections of pathogenic isolates [32]. Taking into account the needs and challenges in population genomics, the following section outlines the prominent approaches for comparing multiple closely related genomes and discusses the opportunities and challenges.

1.3 Comparison of multiple closely related genomes

When sequenced whole genomes became first available, the bioinformatics approaches focused on comparing single proteins or genomic DNA from single genes—the prominent approaches include Needleman and Wunsch [33] and Smith and Waterman [34]. These approaches were variations of edit-distance computations and were ineffective for whole-genome comparisons; the problem statement for a whole-genome alignment incorporating inversions and translocations is NP-complete [35]. This motivated the development of several whole-genome alignment tools, beginning with `MUMmer`, a pairwise whole-genome alignment tool [36]. The increasing availability of sequenced whole genomes mandated the ability to compare several genomes simultaneously. This lead to the development of several approaches (discussed below) for the comparison of multiple whole genomes from related organisms.

1.3.1 Alignment based approaches

1.3.1.1 Whole genome alignment

Whole-genome alignment approaches seek to identify the syntenic chromosomal regions from multiple genomes and align them to support subsequent analysis. Although the problem of finding an optimal multiple sequence alignment is NP-complete [37], tools have been developed that rely on several assumptions and

heuristics to locate putatively homologous regions shared over multiple genomes. Well-known tools include `Mauve` [38], `Mugsy` [39], and `Cactus` [40]. Whole-genome alignment facilitates the detection and analysis of homologous regions, large-scale changes in chromosomes across evolution, discovery of mobile genetic elements, etc. and provides important insights in the structure of genomic rearrangements and horizontal gene transfer events. Whole-genome alignment approaches have a runtime that is quadratic in the number of input genomes. With the increasing availability of sequenced whole genomes, the scalability of these approaches became a major issue [41]. The methods based on whole-genome alignment do not scale adequately for the numbers of genomes (>100) that are relevant and available for current studies.

1.3.1.2 Single Nucleotide Polymorphism based approaches

Single Nucleotide Polymorphism (SNP) based approaches are special cases of conventional whole-genome alignment approaches designed for the identification and analysis of point mutations or SNPs between closely related genomes. They are able to gain speed in performing the genome alignment operations using different assumptions and techniques to determine the 'core' genome and identify SNPs: (i) Core-genome alignment: These approaches focus on identifying the homologous regions conserved from all homologous regions (e.g. `ParSNP` [42]). (ii) Read alignment: These approaches align the reads from the sequenced whole genomes against a reference genome (e.g. `CFSAN SNP Pipeline` [43]). (iii) Exact alignment: These

approaches use k-mers (exact matches of length k) for rapid identification of regions shared between multiple genomes. These k-mers are mapped against a reference genome, similar to that with read alignment, to identify the SNPs (e.g `kSNP` [44]). Although these approaches are extremely fast and scalable compared to conventional whole-genome alignment approaches, they are intrinsically limited to the core genome—which is only about 8% of a typical bacterial genome [45]. Additionally, the reference-guided approaches are biased towards variations present in the chosen reference genomes and suffer in identifying non-reference variations (also referred to as the reference bias), thereby needing closely related reference genomes.

Further, a bacterial species pan-genome could be 'open' or 'closed' [46]: the pan-genome of a species is said to be 'closed' if a limited number of genomes can account for the entire species' genomic diversity, while the pan-genome is said to be 'open' if the gene repertoire keeps increasing under the addition of new genomes. Species with 'open' pan-genomes are poor fit for reference-based approaches.

1.3.2 Gene based approaches

The gene-based approaches seek to predict the genes encoded in given genomes and identify the genes shared across multiple genomes, facilitating a genome comparison via shared gene repertoire. These approaches rely on gene prediction followed by gene clustering to estimate the core and accessory genes, i.e. the genes likely to be present in (nearly) all genomes and those likely to be shared in a subset of the

genomes, respectively. Prominent tools include `PanSeq` [47], `PGAP` [48], and `Roary` [49]. Gene-focused approaches are usually faster than the whole-genome alignment approaches and, unlike the reference-based approaches, they can help in estimating a pan-genome encompassing genes from all given genomes. However, there are issues with both gene prediction as well as gene clustering. Bacterial gene prediction attempts to locate genes from genome sequences by predicting the labels of different regions in genome sequences as protein-coding regions, RNA coding regions, non-coding regions, or intergenic regions [50]; prediction of these regions is adversely impacted both by the assembly quality of the whole genomes and the biases of the gene prediction tool used. Next, the gene clustering step attempts to aggregate the functionally similar gene sequences predicted from multiple genomes into a single gene cluster. Here, we rely on a minimum sequence similarity threshold, say 90%; all gene sequences with at least this sequence similarity are clustered together and are assumed to be functionally equivalent for subsequent analysis. This can, however, create anomalies in genome analysis. For instance, single-nucleotide polymorphisms (SNPs) or insertions-deletions (indels) in a gene can change the gene function. A single SNP can be a determining factor for the resultant phenotype (Figure 1.2) [51]. A sequence similarity based gene clustering may not be able to distinguish the functionally different genes with high sequence identities, and the subsequent analysis using a gene-based pan-genome construction can lead to confounding or misleading inferences [52].

Additionally, gene-based approaches are not whole-genome scale; they are con-

strained to gene boundaries and do not account for intergenic variations. Changes in intergenic regions, such as variations in gene promoter regions, can cause perturbations in gene expression levels and even be the differentiating factors between antimicrobial-resistant and antimicrobial-susceptible isolates. For thorough whole-genome-scale comparisons of multiple genomes, we need to support the analysis of intergenic regions alongside gene-based comparisons. To facilitate such intergenic comparisons, some tools and bioinformatics pipelines have been built over the gene-based approaches and extend the pan-genome to incorporate the intergenic regions [53]. However, as these approaches are an extension of existing gene-based alternatives, they suffer from the same issues related to gene prediction and clustering.

1.3.3 De Bruijn graph based approaches

De Bruijn graph (dBg) represents the overlap information between genome sequences and has been used extensively for genome assembly [54]. The graph consists of a set of nodes or vertices, representing the k-mers (substrings of length k), identified from the genome sequences. An edge connects two vertices if their respective k-mers are such that the $k-1$ length prefix of one k-mer is identical to the $k-1$ length suffix of other k-mer and the corresponding $k+1$ length substring is present in the genome sequence/s. De Bruijn graphs can used for aggregating and representing genomic variations across multiple genomes as well. Here, each genome is assigned a unique 'color' and these colors are assigned to edges to indicate the genome/s where

the edge (i.e., the $k+1$ length substring) is found. The non-branching paths within the graph are 'compacted' to form 'unitigs' (exact matches); the resultant graph is called a compacted de Bruijn graph (cdBg). Well-known tools for compacted de Bruijn graph representation of multiple genomes include `SplitMEM` [55], `TwoPaCo` [56], and `Cuttlefish` [57]. De Bruijn graph approaches are much faster and scale to a large number of genomes compared to whole-genome alignment approaches. However, a compacted dBg from, say 50, bacterial genomes results in hundreds of thousands of unitigs. A comparative genome analysis using de Bruijn graph suffers from a 'short-fat matrix' formulation—hundreds of thousands of genomic features (unitigs and/or edges) to be analyzed from a countably few genomes (hundreds or a few thousand)—making the downstream assessments extremely challenging.

1.3.4 Specialized approaches for genotype-phenotype analysis

The above-mentioned approaches for comparison of multiple closely-related genomes have been limited in their utility for genotype-phenotype analysis, either due to scalability constraints, challenges due to feature dimensionality (short-fat matrix formulation), or limitations of the gene-based approaches. Owing to these limitations, some alternative strategies have been developed to cater to specific genotype-phenotype correlation studies. For instance, in the case of antimicrobial resistance, many analyses rely on the detection of known resistance genes. These genes are curated and compiled into AMR gene databases, such as CARD [58],

ARDB [59], ARG-ANNOT [60]. Some tools also use these AMR gene databases to train machine learning models and predict the AMR phenotypes of given isolates, e.g. DeepARG [61], PATRIC [62], etc. The efficacy and reliability of these tools and approaches heavily rely on the knowledge about the known AMR factors; these approaches cannot check for novel resistance factors.

Bacterial genome-wide association studies (GWAS) are some alternative approaches developed. GWAS have been extensively used in human populations for case-control studies: individuals with a disease or condition under study form the 'cases' while a similar group of people who do not have the condition form the 'controls'. In human GWAS, SNPs have been used to identify the loci associated with complex phenotypes and facilitate genetic discoveries linked to variety of diseases [63, 64]. Bacterial GWAS employ alignment-free comparison strategies: they extract k-mers from assembled genomes or their read sequences and model k-mer distributions which can be used to predict the phenotype. Prominent tools for bacterial GWAS include `SEER` [65] and `DBGWAS` [66]. The number of distinct k-mers present in genomes largely exceeds the genome counts (much more than the SNPs in human GWAS)—a subsequent analysis would suffer from feature dimensionality issues. Therefore, to discern the k-mers corresponding to relevant genomic features, these approaches have to rely on various heuristics to constraint the k-mers to be evaluated. Additionally, the alignment-free comparisons with extracted k-mers loses the genomic context in these analyses. These analyses are oblivious to the genomic rearrangements which may influence the phenotypic variations. For instance, pro-

moter region changes are known to alter the gene expressions and form an important AMR mechanism (Fig. 1.2). Overall, the applications of specialized approaches for genotype-phenotype analysis are limited to certain specific cases and preclude a general-purpose whole-genome-scale population genomic exploration.

Apart from the shortcomings of genome comparison approaches, bacterial population genomics largely relies on the analysis of fragmented draft assemblies—a collection of contigs (fragmented contiguous DNA sequences) constitutes an isolate's genome. Fragmentation of genome assemblies, however, presents several limitations and challenges for comparative genomics [67, 68]. For instance, different mobile genetic elements, transposable elements, genomic repeats, and clusters of horizontally transferred genes—commonly referred to as the genomic islands—are known to adversely impact the assembly quality leading to fragmented assemblies [68, 69, 70]. These genomic islands and mobile genetic elements are often assembled into separate contigs in draft assemblies, and can include several AMR and virulence factors in pathogenic microbes [21, 69, 70]. Some mobile genetic elements encode insertion sequences and can act as gene promoters that upregulate certain AMR genes, enabling the isolates to confer antimicrobial resistance [21, 71]. An incomplete, fragmented genome constrains our ability to ascertain if such genomic regions, like these insertion sequences, are in the specific places, creating ambiguities in the genotype-phenotype association studies.

In principle, to support a robust and thorough large-scale population genomic, we would want an approach that can handle a large number of genomes and facilitate

a downstream analysis of a concise collection of genomic regions shared between multiple isolates alongside the genomic context of these regions. With this in mind, we developed our novel approach, `PRAWNS`, that generates compact pan-genomic features for population genomics, described in detail in Chapter 4.

1.4 Quasimetagenomics

Epidemiological monitoring and spread of infectious diseases have historically relied on case counts from clinical diagnosis [72]. The need for early detection and monitoring the spread of pathogens lead to the development of early foodborne disease programs, such as `PulseNet`, which performed routine surveillance of bacterial pathogens using molecular subtyping by pulsed-field gel electrophoresis [13]. The application of whole-genome sequencing has revolutionized strain-typing and source attribution of bacterial pathogens; it has facilitated faster and cost-effective surveillance of pathogens, thereby transforming public health and food safety monitoring [13, 14, 73]. A salient objective for pathogen detection and monitoring in public health settings is pathogen source tracking—identifying the source of microbial contamination [74]. For this, several food, clinical, veterinary, or environmental samples are collected to check for the presence of pathogens. Pathogens, however, occur at low abundances in such microbial communities and may not provide sufficient DNA for sequencing and efficient analysis. Current approaches rely on selective culture enrichment—the microbial community is clinically cultured or grown in a selective

medium which 'selects' a particular microorganism for growth in that medium to provide a pure-culture colony of microbial isolates. Selective enrichment and pure colony isolation of target pathogens form the preparatory steps of these approaches for sequencing and analysis. These steps are labor-intensive and can take days, even weeks, to get sufficient DNA for sequencing [75, 76, 77]. To expedite this process, a technique called quasimetagenomic sequencing has been proposed [78]. Quasimetagenomics refers to sequencing of samples that have not undergone complete enrichment to form pure culture bacterial isolates, or abbreviated recovery enrichments from the microbial samples [79, 80]. The quasimetagenomic approach makes it possible to perform outbreak monitoring at a much earlier time point after the sample collection.

Although quasimetagenomics promises expedited sequencing and data availability for analysis, the sequenced samples are not pure-culture isolates; existing approaches for genome assembly and analysis may not be well-suited for quasimetagenomics. The abbreviated recovery enrichments may include multiple strains of a species and isolates from different species; the sequenced DNA can comprise a variety of genomic repeats arising from both inter- and intra-species repeat regions, which can affect the assembly quality and subsequent bioinformatics assessments. It is, therefore, unclear whether the existing approaches (with any parameter tunings) would be useful for quasimetagenomics. More benchmarking is required to understand the effectiveness of existing approaches, avenues for novel developments, and formulations of new pathogen monitoring standards using quasimetagenomics.

1.5 Metagenomics

Metagenomics refers to the DNA extraction and sequencing from a microbial community directly from its environment. It has proven to be a key tool for the analysis of complex microbial communities, particularly the member organisms that cannot be cultured in the laboratory using conventional isolation [81, 82]. Metagenomics has facilitated the characterization of microbial diversity in the environment, identification of novel enzymes, bioprocesses, and natural products—such characterizations enable the assessment of potential microbial applications in various industrial and biomedical settings, such as development of novel antibiotics, biocatalysts, industrial enzymes etc. [81, 83, 84]. Some of the applications where metagenomics has been increasingly adopted include food safety monitoring, spread of antimicrobial resistance factors, public health surveillance, and other comparative microbiome studies [85, 86, 87].

Metagenomics can facilitate case-control microbiome studies: we can compare microbial compositions between healthy individuals (controls) and diseased subjects or patients (cases). These studies define objects (genomic entities) that can be recognized and measured across samples—such as SNPs, k-mers, genomes, taxonomic labels, and genes—and associate their abundance or presence with phenotypes [88, 89]. Owing to challenges in accurate identification of these objects from metagenomic samples and feature dimensionality issues (millions of potential metagenomic features possible from a countably few samples), case-control micro-

biome studies have primarily relied on taxonomic identification and associating taxa presence-absence and relative abundances with cases or controls [85, 90]. However, the mere presence of a taxon is not always a deterministic factor for disease causality [91, 92, 93]. For instance, *E. coli* could be present in metagenomic samples collected from both healthy individuals and patients; the presence of Shiga-toxin producing isolate of *E. coli* (STEC), however, is more likely to be present in patient samples. It is, therefore, desirable to perform case-control microbiome analysis at a whole-metagenome-scale and explore the strain-level variations, structural variations and genomic regions corresponding to unclassified or understudied microbes.

The advances in next-generation sequencing (NGS) technologies have contributed to unprecedented amounts of shotgun metagenomic sequencing data; publicly available repositories, like NCBI, currently host order of magnitude more shotgun metagenomic data than what they did ten years ago [94]. These public repositores have presented opportunities to perform whole-metagenome-scale microbial diversity analyses using samples from large and diverse populations (even across geographical locations or from different time-points) and contrast the findings from previous studies on smaller metagenomic cohorts. These analyses could also help in case-control microbiome studies to uncover the genomic similarities and differences between case populations infected with a disease contrasted against healthy controls incorporated from public datasets [95].

Conceptually, the comparisons between multiple metagenomic samples draw parallels with the genomic comparisons of bacterial isolates. In both instances, we

get sequencing reads which are typically assembled to construct assembled genomes or metagenomes for the pure-culture isolates or microbial communities, respectively. These assemblies are compared to explore the genomic regions shared between multiple sequenced samples. However, metagenomic sample comparisons are far more challenging: each metagenomic sample can have a unique microbial diversity (i.e., microbes present uniquely in individual samples) and also contain several unknown or understudied microbes [96], which, in combination, make the identification of genomic regions shared between multiple metagenomic samples more complicated compared to that for pure-cultured isolates. Existing approaches for metagenomic comparisons either rely on 'co-assembly' of metagenomic samples (i.e., metagenomic assembly using sequencing reads from all metagenomic samples under study) or aligning the metagenomic reads against curated sequence databases [97, 98]. The co-assembly approach requires shotgun metagenomic reads from all samples to be assembled together and genomic variations within individual samples are later identified by via read alignment. This approach is computationally expensive and does not scale beyond a handful of samples. Read alignment to reference sequences scales to more than several samples but is limited by the biases in reference-based comparisons—a bias towards identification of genomic variations present in the reference genomes and the need for reference genomes that are closely-related to the microbial community under study. The existing approaches are, hence, unable to facilitate a reference-free whole metagenomic exploration and analysis that scales to large collections (hundreds or thousands) of samples while supporting the assessment of genomic regions from unclassified or understudied taxa. To fill this gap, we

developed SIMILE, a novel approach for alignment-free extraction of similar genomic regions shared between multiple metagenomic samples, described in Chapter 6.

1.6 Contributions

In Chapters 2 and 3, I present two of my works on identifying genomic determinants of antimicrobial resistance using population genomics [20, 21]. Chapter 2 describes my contributions using large-scale whole-genome assessment of *Acinetobacter baumannii* isolates to identify the genomic factors associated with carbapenem resistance (one of the last-line antibiotics) [21]. We identify the previously known and strongly correlated AMR factors as well as putative novel factors contributing to carbapenem resistance. In Chapter 3, we use population genomics to track the world-wide spread of a plasmid that encodes AMR factors in multi-drug resistant (MDR) *Shigella* [20]. Our work on *Shigella* explores connections between the multidrug resistance phenotype of isolates and the presence of a plasmid.

In Chapter 4, we present PRAWNS: fast and scalable tool that generates a compact collection of pan-genomic features for large number of closely related whole genomes. Chapter 4 describes the method and its utility for population genomics analysis. PRAWNS enables the identification of conserved regions and collocated regions shared between multiple genomes. It supports contig scaffolding information, thereby making is possible to integrate the assessment of fragmented contigs for a thorough analysis.

In Chapter 5, we explore the utility of quasimetagenomics and the integration of long and short-read sequencing technologies to expedite the accurate reconstruction of a pathogen genome. Our work evaluates the accuracy of *Listeria monocytogenes* assemblies constructed from quasimetagenomics samples; the samples correspond to a contaminated ice cream that caused a foodborne illness outbreak [99]. Here, we assess the efficacy of using quasimetagenomics when contrasted against pure culture isolates, benchmarking ten prominent approaches for short-read, long-read, and hybrid assembly, and identify the biases of these assembly approaches and areas for future research.

In Chapter 6, we present `SIMILE`, a framework for alignment-free extraction of similar genomic regions shared between multiple assembled metagenomes. The input to `SIMILE` is individually assembled metagenomes. `SIMILE` scales to hundreds of metagenomic assemblies and facilitates the rapid discovery of resembling genomic regions (contigs) shared across the metagenomic samples. The results highlight that these whole-metagenome-scale comparisons using `SIMILE` can leverage microbial diversity analyses to scale at population metagenomics or large-scale comparative metagenomics and expedite microbiome research.

To summarize my contributions from my PhD research, the analyses of several samples can facilitate better insights into microbial genome biology. Large-scale comparisons have the ability to assess a wide range of genomic variations and also evaluate their influence on various outcome scenarios. Microbial genomics continues to suffer from the 'large p, small n' (short-fat matrix) problem and demands

the use of techniques which can disambiguate the 'signal' from 'noise'. Through my research, I have emphasized the types of genomic features that are biologically relevant and developed bioinformatics tools which can both scale to large genomic or metagenomic datasets and also enable the identification of biologically relevant features.

Chapter 2: Carbapenem Resistance in *Acinetobacter baumannii*

This chapter contains material previously published in Whole-Genome Assessment of Clinical Acinetobacter baumannii Isolates Uncovers Potentially Novel Factors Influencing Carbapenem Resistance [21], which was a joint work with Hugh Rand, Maria Hoffmann, Yan Luo, Saul Sarria, Nagarajan Thirunavukkarasu, Christine A. Pillai, Patrick McGann, J. Kristie Johnson, Errol Strain and Mihai Pop. KJ performed the literature review, most of the short- and long-read assemblies, all bioinformatic analyses, wrote most of the manuscript with contributions from all authors, and created the figures. NT and CP performed the microbiological and molecular lab work. MH and YL performed the PacBio-SMRT long read assemblies. SS performed the carbapenem phenotype testing. JKJ provided the carbapenem susceptibilities for the 349 A. baumannii isolates. KJ, HR, PM, ES, and MP edited the manuscript.

2.1 Introduction

Antimicrobial resistance (AMR) is a grave healthcare challenge worldwide, causing 700,000 deaths each year [100]. A sizable proportion of AMR bacterial infections are hospital-acquired and pose a high healthcare burden in developed and developing countries alike [27]. Some of these bacterial infections can be resistant to even the most powerful antibiotics available; including carbapenems which are typically considered to be one of the "last-line" antibiotics by clinicians, and specifically used in the treatment of critically-ill patients affected potentially by antimicrobial-resistant Gram-negative infections [28]. Antimicrobial resistance is caused by a wide variety of genetic factors across different pathogens; discovering and understanding these factors has been accelerated by the advancements in whole genome sequencing technologies and the development of methods for analyzing the resulting data [1].

Acinetobacter baumannii, a Gram-negative and frequently multidrug-resistant pathogen, is an important nosocomial pathogen worldwide [101]. The Infectious Diseases Society of America considers *A. baumannii* to be among the top six leading causes of nosocomial infections, i.e., one of the high-priority ESKAPE pathogens [102]. The World Health Organization has marked carbapenem-resistant *Acinetobacter baumannii* (CRAB) as their highest priority pathogen for the research and development of novel antibiotics [103]. *A. baumannii* has the ability to proliferate readily in hospitals, particularly in intensive care units, and spread epidemically among patients [27]. It can cause a broad range of severe infections, such as

skin and soft tissue infections, wound infections, urinary tract infections, secondary meningitis, and ventilator-associated pneumonia [104, 105]. Its ability to have intrinsic resistance and propensity to acquire resistance has led to the emergence of multidrug-resistant, extensively drug-resistant, and pan-drug resistant *A. baumannii* strains [71, 106]. An increasing rate of carbapenem resistance was documented among *A. baumannii* strains from numerous hospital outbreaks [107, 108].

A. baumannii strains have the ability to resist the action of carbapenems via both intrinsic properties and acquired resistance factors [71]. Carbapenem resistance is primarily mediated by the production of carbapenem-hydrolyzing beta-lactamases, also called carbapenemases [104, 109, 110]. Beta-lactamases have been historically categorized into four molecular classes (A to D) based on conserved amino-acid motifs. Carbapenem resistance in *A. baumannii* is largely attributed to the class D beta-lactamases, also known as OXA-type carbapenemases [111]. The other major mechanism conferring carbapenem resistance in *A. baumannii* is the modification of membrane permeability, either by the loss or decrease in the expression of outer membrane proteins, increased expression of efflux pumps, or modifications in penicillin-binding proteins (PBPs) [112]. The known factors that influence carbapenem resistance in *A. baumannii* are described in Table 2.1.

CRAB strains are also resistant to a number of non-carbapenem antibiotics. Mechanisms of resistance include aminoglycoside-modifying enzymes, sulfonamide resistant genes, quaternary amines, other resistance genes encoded within the AbaR resistance islands [110, 133]. Similarly, CRAB strains may encode transcriptional

Genetic element	**Mechanism for carbapenem resistance**	**Literature supporting gene-resistance link**	**Literature indicating insufficient support**
Carbapenem hydrolysis by beta-lactamase genes			
Intrinsic beta-lactamases (*OXA-51-like*, *ADC*, etc.)	Weak hydrolysis Primarily chromosomal	[113, 114, 115]	[116, 117, 118]
Acquired beta-lactamases (*OXA-23*, *OXA-58*, etc.)	Strong hydrolysis Primarily plasmid-encoded	[110, 116, 119]	[120, 121, 122]
Insertion Sequences (IS*Aba1*, IS*Aba3*, etc.)	Upregulate the expression of beta-lactamase genes when present in their promoter region	[111, 118, 123]	[110, 124]
Loss of membrane permeability			
Outer membrane proteins (*ompA*, *carO*, etc.)	Mediated by loss or inactivation of genes	[125, 126]	[127]
Penicillin-binding proteins (PBPs)	Mediated by modification or loss of PBPs	[128, 129]	[123]
Efflux pumps (*adeABC*, *adeIJK*, etc.)	Over-expression of efflux pumps	[119, 130]	[131, 132]

Table 2.1: Genomic drivers known to influence carbapenem resistance in *Acinetobacter baumannii*.

regulator and transporter genes which are known to expedite resistance to non-carbapenem antibiotics by regulating the expressions of influx/efflux pumps, biofilm formation, outer membrane permeability, etc. [131, 134] The impact of these elements on carbapenem resistance has not been fully quantified, and they do occur in carbapenem-susceptible isolates as well [135, 136]. Overall, the resistance phenotype of an isolate is likely a product of the interactions of different genomic factors, and we currently lack a full understanding of the complete set of such genomic factors and their interactions. Here, we utilize a large-scale whole-genome comparison to address some gaps in our understanding of carbapenem resistance in *A. baumannii*.

Our work expands upon the findings of a recent study of 203 *A. baumannii* isolates collected as a part of an active infection control surveillance program at the University of Maryland Medical Center (UMMC) [122]. We increase the number of genomes analyzed to 349 (refer Supplementary Table 1 from [21] for NCBI Accessions) and expand the analysis beyond individual genes and conventional genome-wide association studies (GWAS). The analysis of the association between genotype and phenotype encompasses known resistance genes, insertion sequences, plasmids, single nucleotide polymorphisms (SNPs), and other conserved regions. Of particular interest is a subset of genetically similar strains for which the resistance phenotype could not be explained by the known determinants of carbapenem resistance. In this subset, we employ sophisticated whole genome analysis approaches and identify several genomic features that are associated with antibiotic resistance within these isolates—providing several intriguing opportunities for future research.

2.2 Results

2.2.1 Genomic Data, Antimicrobial Susceptibility Testing, and MLST

The 349 *A. baumannii* isolates were collected in clinical setting as a part of a surveillance program at the University of Maryland Medical Center (UMMC); these isolates had been sequenced and are accessible through NCBI (Supplementary Table 1 from [21]). The level of resistance to several carbapenems was obtained using the

Kirby-Bauer disk diffusion method, and phenotypes were determined according to the Clinical Laboratory Standards Institute (CLSI) breakpoints. Here, we restrict our analysis to imipenem, for which information was available for all isolates (Supplementary Table 1 from [21]). Out of the 349 isolates, 305 were imipenem-resistant and 44 were imipenem-susceptible. Since the sequence assemblies available at NCBI were highly fragmented (mean contig count: 145) and had a higher number of misassemblies (mean misassemblies count: 84, mean mismatches per 100 kbp count: 1,003), we re-assembled the sequencing reads for each isolate (mean contig count: 126, mean misassemblies count: 59, mean mismatches per 100 kbp count: 729) and re-annotated the resulting assemblies, as described in Methods.

MLST sequence types (ST) were determined for the isolates using the Pasteur scheme [137]; the resultant MLST profiles are presented in Supplementary Table 1 from [21]. Two hundred and thirty-six out of the three hundred and forty-six isolates were assigned to the Pasteur ST 2, which was also the largest ST category identified in Wallace *et al.* study [122].

2.2.2 Presence of Known Antimicrobial Resistance Genes

Known antimicrobial resistance (AMR) genes were found in all isolates (Supplementary Table 2 from [21]). Some of the known AMR genes exhibited a strong association with the imipenem-resistant phenotype ($p \leq 0.01$, Fisher's exact test), but were also found in many imipenem-susceptible isolates as well (Supplementary

Table 2 from [21]), indicating that alone those genes do not explain resistance. To identify genes that are most likely to confer resistance in the absence of other factors, we focused on genes with strong statistical association with the imipenem-resistance phenotype ($p \leq 0.01$) and which were infrequently found in imipenem-susceptible isolates (≤ 10 out of 44). Ten AMR genes were strongly correlated with the imipenem-resistant phenotype according to these criteria (Table 2.2 and Supplementary Table 1 from [21]). One or more of these genes was identified in 294 (84%) *A. baumannii* isolates (288 imipenem-resistant and 6 imipenem-susceptible), suggesting that for the majority of the isolates (82%) the imipenem-resistant phenotype can be explained by their presence.

Gene annotation	**Gene present**		**Gene absent**		**Odds Ratio**
	Resistant	**Susceptible**	**Resistant**	**Susceptible**	
bla$_{\text{OXA-23}}$	182 (60%)	1 (2%)	123 (40%)	43 (98%)	63.62
qacEdelta1	257 (84%)	4 (9%)	48 (16%)	40 (91%)	53.54
sul1	259 (85%)	5 (11%)	46 (15%)	29 (89%)	43.91
mphE	175 (57%)	2 (5%)	130 (43%)	42 (95%)	28.27
msrE	174 (57%)	2 (5%)	131 (43%)	42 (95%)	27.89
ant(3")-II	158 (52%)	2 (5%)	147 (48%)	42 (95%)	22.57
aacC1	156 (51%)	1 (2%)	149 (49%)	43 (98%)	45.02
yafP	155 (51%)	1 (2%)	150 (49%)	43 (98%)	44.43
aphA6	74 (24%)	1 (2%)	231 (76%)	43 (98%)	13.77
xerD	72 (24%)	1 (2%)	233 (76%)	43 (98%)	13.29
Combined total	288 (94%)	6 (14%)	17 (6%)	38 (86%)	107.29

Table 2.2: Ten known antimicrobial resistance genes strongly associated with imipenem-resistance in 349 *A. baumannii* genomes analyzed ($p \leq 0.01$, Fisher's exact test) and present infrequently in imipenem-susceptible isolates (≤ 10 out of 44).

Wallace *et al.*'s analysis of 203 *A. baumannii* isolates identified eight genes not previously considered to be AMR genes that were strongly associated with carbapenem resistance [122]. Within the larger set of 349 isolates (that includes the

203 strains analyzed by Wallace *et al.*), these genes continued to be significantly-associated with resistance (Table 2.3). Seven of these eight genes were detected exclusively in imipenem-resistant isolates—each of these isolates also contained one or more of the 10 strongly correlated AMR genes (Table 2.2).

Gene ID	**Annotation**	**Gene present**		**Gene absent**	
		R	**S**	**R**	**S**
55125	Polysaccharide biosynthesis family protein	101 (33%)	0 (0%)	204 (67%)	44 (100%)
74448	Glycosyl transferases group 1 family protein	102 (33%)	0 (0%)	203 (67%)	44 (100%)
197234	KR domain protein	102 (33%)	0 (0%)	203 (67%)	44 (100%)
197235	Cytidylyltransferase family protein	102 (33%)	0 (0%)	203 (67%)	44 (100%)
365901	Type IV pilin structural subunit	51 (17%)	18 (41%)	254 (83%)	26 (59%)
563378	Conserved hypothetical protein	102 (33%)	0 (0%)	203 (67%)	44 (100%)
733161	Oxidoreductase, NAD-binding Rossmann fold family protein	101 (33%)	0 (0%)	204 (67%)	44 (100%)
733163	Capsule polysaccharide biosynthesis family protein	101 (33%)	0 (0%)	204 (67%)	44 (100%)

Table 2.3: Prevalence within our collection of isolates of genes found by Wallace *et al.* to be associated with imipenem-resistance in *A. baumannii*. (R) and (S) denote the imipenem-resistance or susceptible isolates, respectively.

2.2.3 Impact of Insertion Sequence IS*Aba1* on Resistance

Two beta-lactamase genes known to elevate resistance to carbapenems in *A. baumannii*, $bla_{\text{OXA-65-like}}$ ($bla_{\text{OXA-51}}$ family) and bla_{ADC}, were found in all 349 isolates analyzed in our study, indicating their presence alone is not sufficient for resistance.

These two genes in conjunction with an upstream insertion of a mobile genetic element, the IS*Aba1* insertion sequence, have been previously associated with resistance (Table 2.1). However, the analysis of insertion sequences and other mobile elements is complicated due to the challenges they pose to assembly algorithms [70, 138], a fact recapitulated in our data where IS*Aba1* was identified predominantly on a separate genomic contig. To provide a genomic context for this sequence with respect to other resistance genes, we relied on scaffolding information based on paired-reads. This information allowed us to ascertain the relative distance between IS*Aba1* and resistance genes (see Supplementary Table 3 from [21]). The IS*Aba1* sequence was detected in 317 of the 349 isolates, where 296 are resistant and 21 are susceptible. As shown in Table 2.4, the presence of IS*Aba1* in the upstream region of the $bla_{\text{OXA-65-like}}$ and bla_{ADC} genes was strongly associated with resistance ($p \leq 0.01$, Fisher's exact test), and we also recapitulated the strong association between IS*Aba1* and $bla_{\text{OXA-23}}$.

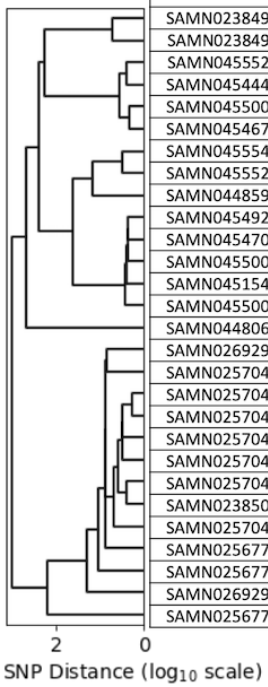
Gene Annotation	**Both, IS*Aba1* and gene, present**				**IS*Aba1* absent, gene present**	
	IS*Aba1* is upstream of the gene		**IS*Aba1* is not upstream of the gene**			
	R	**S**	**R**	**S**	**R**	**S**
$bla_{\text{OXA-23}}$	110 (36%)	0 (0%)	72 (24%)	1 (2%)	0 (0%)	0 (0%)
$bla_{\text{OXA-51}}$	148 (49%)	0 (0%)	148 (49%)	21 (48%)	9 (2%)	23 (52%)
bla_{ADC}	145 (48%)	2 (5%)	151 (50%)	19 (43%)	9 (2%)	23 (52%)

Table 2.4: Imipenem resistance is potentiated by the presence of the IS*Aba1* insertion sequence upstream of resistance genes. (R) and (S) denote the imipenem-resistance or susceptible isolates, respectively.

2.2.4 Microbial Clade Lacking Known Antimicrobial Resistance Genes

Fifty-five isolates (17 imipenem-resistant and 38 imipenem-susceptible) did not encode either of the 10 genes we have found to be strongly associated with imipenem-resistance. The imipenem resistance within the resistant isolates could not be directly explained by the known AMR genes, or the presence of the IS*Aba1* insertion sequence in the upstream region of the beta-lactamase genes. We estimated the phylogeny of all 55 isolates based on the single nucleotide polymorphism (SNP) matrix computed with the reference-based `CFSAN SNP Pipeline` (Supplementary Figure 1 from [21]) [43]. This analysis revealed a clade that comprised 15 isolates: 13 of the 17 imipenem-resistant isolates and 2 imipenem-susceptible isolates. All isolates from this clade displayed similar known AMR gene profiles (Figure 2.1). In order to better explore the genomic differences between the imipenem-resistant and susceptible isolates from this clade, an additional 13 sequences of imipenem-susceptible strains were retrieved from the NCBI Pathogen Detection Isolates Browser. These *A. baumannii* isolate sequences were located in the same SNP cluster as the genome sequences from the above-mentioned clade (cluster ID: PDS000005681). The NCBI identifiers for these 13 isolates along with their carbapenem susceptibilities and phenotypes are provided in Supplementary Table 4 from [21]. As we have done for all isolates discussed in this paper, their sequences were re-assembled and subjected to gene prediction and annotation. Furthermore, we were able to obtain biological samples from 12 of these isolates and confirm the phenotype reported in the NCBI

database (susceptible to imipenem, Supplementary Table 4 from [21]).



BioSample	Phenotype	AST	Beta-lactamases			Aminoglycoside-modifying enzymes					
			OXA-65	*ADC*	*TEM-1*	*ANT(3")-II*	*APH(3")-Ib*	*APH(6)-Id*	*AacA16*	*sul1*	*qacEdelta1*
SAMN02384922	Susceptible	22 (DD)	✓	✓		✓	✓	✓		✓	
SAMN02384921	Susceptible	23 (DD)	✓	✓		✓	✓	✓		✓	
SAMN04555295	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04544432	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04550089	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04546730	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04555488	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04555209	Susceptible	2 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04485911	Susceptible	2 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04549259	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04547094	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04550081	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04515409	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04550084	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN04480665	Susceptible	1 (MIC)	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAMN02692995	Resistant	17 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02570436	Resistant	17 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02570433	Resistant	17 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02570432	Resistant	17 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02570431	Resistant	17 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02570435	Resistant	17 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02570434	Resistant	17 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02385032	Resistant	17 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02570430	Resistant	17 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02567743	Resistant	16 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02567735	Resistant	16 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02692975	Resistant	14 (DD)	✓	✓		✓	✓	✓	✓	✓	
SAMN02567753	Resistant	16 (DD)	✓	✓		✓	✓	✓	✓	✓	

Figure 2.1: The lineage estimated for the chosen 28 isolates (y-axis) based on SNP distance (x-axis (log10 scale)). The Antimicrobial Susceptibility Testing (AST) values provide the imipenem susceptibilities for the corresponding isolates measured via disk diffusion (DD) or minimum inhibitory concentration (MIC). None of these isolates were predicted to contain any of the 10 strongly correlated known AMR genes presented in Table 2.2, but contain several other known AMR genes. The imipenem-resistant and susceptible isolates cluster separately.

To assess whether known determinants of resistance, beyond the 10 genes described above, could determine the phenotype of these 28 isolates (15 from our study plus 13 retrieved from NCBI), we predicted their phenotype using `PATRIC` (v3.6.5 online) [62]. The computational predictions disagreed with the experimentally-determined phenotypes: all 13 imipenem-resistant isolates were predicted by `PATRIC` to be susceptible, while 3 of the imipenem-susceptible isolates were predicted to be resistant (Supplementary Table 5 from [21]).

2.2.5 Pan-Genome Analysis for Genomic Variants Associated With Resistance Phenotypes

The 28 isolates had high overlap in gene content—among the 4,092 genes within their collective pan-genome, 3,316 (81%) constituted the core genes (genes found in at least 99% of these genomes). Among the core genes were several known AMR genes and virulence factors (Supplementary Table 6 from [21]). All 28 isolates were assigned to the Pasteur ST 2 MLST profile. Despite the high genomic similarity between these isolates, the imipenem-resistant isolates clustered within a clade distinct from that of the imipenem-susceptible ones (Figure 2.1). The insertion sequence IS*Aba1* was detected upstream of the beta-lactamase genes $bla_{\text{OXA-65-like}}$ as well as bla_{ADC} in imipenem-resistant as well as imipenem-susceptible isolates (Supplementary Table 6 from [21]). The Antimicrobial Susceptibility Testing (AST) values for these 28 isolates (Figure 2.1) were extremely close to the CLSI breakpoints that determined the respective imipenem-resistance phenotypes-suggesting the presence of genomic variants that influence the imipenem susceptibilities just enough to alter the phenotype in the absence of strongly associated imipenem-resistance genes.

We performed a whole-genome comparison of the isolates to identify genomic regions that are present exclusively in resistant or susceptible isolates. To improve our chances of detecting contiguous genomic segments that may contribute to imipenem-resistance, we re-sequenced 6 of the isolates using the long read Pacific Biosciences (PacBio) technology, obtaining complete genome sequences for these iso-

lates (Supplementary Table 4 from [21]). This analysis revealed a group of transcriptional regulators and transporters that were found exclusively within the imipenem-susceptible isolates. The transcriptional regulators were predominantly helix-turn-helix (HTH) type DNA binding transcriptional regulators, such as *gntR* family, *iclR* family, *lysR* family, *marR* family, and *tetR* family. These genes were encoded in a 38,651 nt contiguous chromosomal region (Figure 2.2)—this region is conserved across all 15 imipenem-susceptible isolates and absent from all 13 imipenem-resistant isolates (Supplementary Table 5 from [21]).

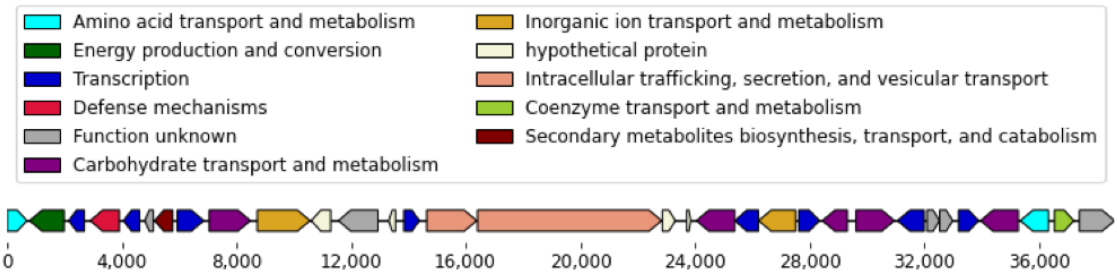


Figure 2.2: Genes encoded in the 38,651 nt chromosomal gene cassette. Within the chosen 28 isolates, this chromosomal gene cassette was conserved in all 15 imipenem-susceptible isolates and absent from all 13 imipenem-resistant isolates.

Whole-genome resequencing of the isolates also allowed us to identify two plasmids with mutually-exclusive presence (Supplementary Table 5 from [21]). The longer plasmid is 111 kbp in length and is nearly identical to the *Acinetobacter baumannii* plasmid *pAYP-A2* [129] (100% query coverage, 99.99% identity). This plasmid sequence was identified in 14 of the 15 imipenem-susceptible isolates and was missing from the remaining 14 isolates. This plasmid contained a number of genes identified by Wallace *et al.* [122] to be associated with susceptibility to imipenem (Supplementary Table 2 from [21]). All 14 isolates that lacked this plasmid

contained another plasmid of length 14.6 kbp length. The shorter plasmid was most similar to an *A. baumannii* plasmid *pORAB01-3* with 75% query coverage and 99.9% identity [139]. Neither plasmid encodes any of the known AMR genes. Note that the previously sequenced *pORAB01-3* plasmid contains an *OXA-134* family class D beta-lactamase gene *OXA-237* flanked on either side by IS*Aba1* elements, feature absent from the corresponding plasmid identified in our isolates.

In addition to the 38.6 kbp chromosomal gene cassette and the two plasmids, the whole-genome comparisons also allowed us to locate a number of genomic variants that were strongly associated with the imipenem-resistant or susceptible isolates (p-value ≤ 0.01, Fisher's exact test). These genomic variants included both SNPs and longer genomic segments of up to thousands of nucleotides in length. These include 75 distinct conserved regions that were at least 500 nt in length. Fourteen of these regions were also found in the reference *A. baumannii* strain [Genbank ID: NC_017162] along with 254 SNPs (Supplementary Table 8 from [21]). These features were distributed fairly evenly across the reference genome (Figure 2.3).

2.3 Discussion

Previous studies on the association of genetic elements with carbapenem resistance in *A. baumannii* have focused on specific genes or were confined to small numbers of isolates [109, 140, 141]. Our study on imipenem, a type of carbapenem antibiotic, resistance is broader and provides a whole-genome analysis of 349 iso-

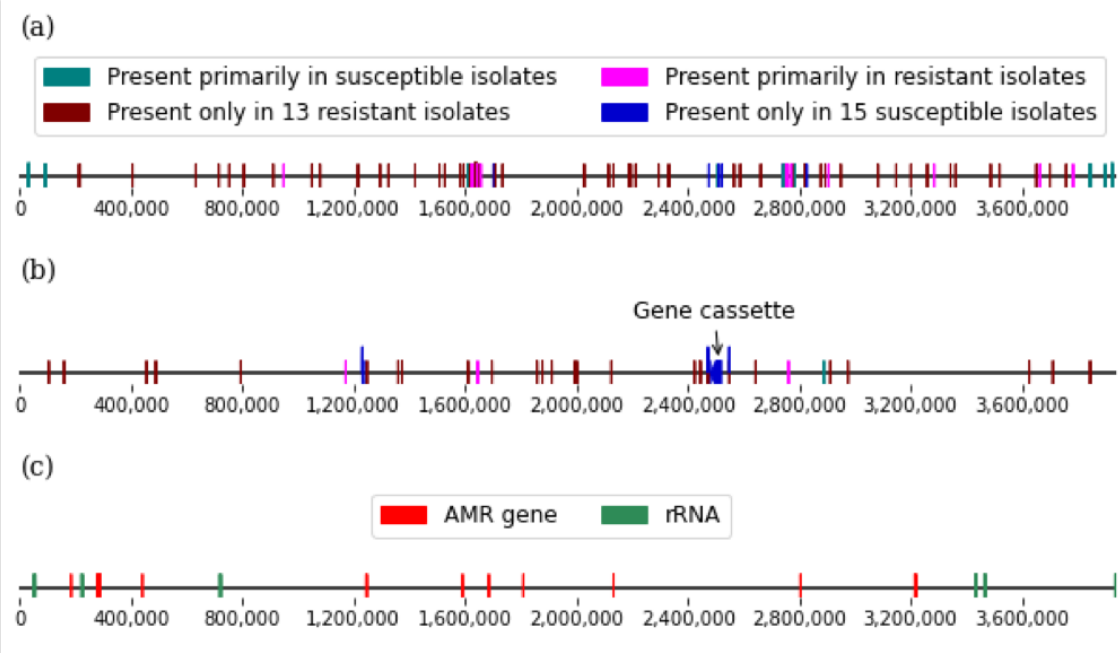


Figure 2.3: (a) Single nucleotide polymorphism (SNP) and (b) conserved genomic regions ($\geq$ 500 nt in length) strongly associated with the imipenem-resistance phenotypes and located on the reference genome. (c) Locations of the known resistance genes and ribosomal RNA genes in the reference genome.

lates. The whole-genome approach we employed included an analysis of pairs of genomic factors; an important consideration for factors such as IS*Aba1* which have the potential to affect the expression of neighboring genes. Our approach, enabled by improved genome assemblies and genome scaffolding information, also allowed the assignment of factors to either chromosomes or plasmids, providing a better understanding of the potential for mobilization of the genomic factors identified.

Our analysis identified ten genes that were strongly associated with imipenem resistance in the *A. baumannii* isolates we analyzed, and confirmed the association with resistance of genes identified by Wallace *et al.* when analyzing a subset of the genomes we explore here. Among the ten genes we have found to be strongly associated with resistance, only $bla_{\text{OXA-23}}$ is widely believed to be involved in car-

bapenem resistance [116]; the other genes are associated with resistance to other antibiotics [58, 142, 143], or are involved in the mobilization of resistance genes. This highlights a limitation of computational-only methods such as ours—causality can only be demonstrated experimentally. The results presented here provide several promising targets for experimental validation.

An intriguing finding of our study is that several genomic regions seem to be associated primarily with susceptibility to carbapenems even in the presence of known AMR genes. These genomic regions appear to contain a high density of transcriptional regulators, suggesting a potential role of these genes in the etiology of antimicrobial resistance in *A. baumannii*. Supporting the plasticity of these plasmids is the fact that a previously-sequenced variant of one of the plasmids we identified, *pORAB01-3*, contains an antimicrobial resistance gene that was not present in the genomes we analyzed. Given that the actual AST values of the chosen imipenem-resistant and susceptible isolates were close to the CLSI breakpoints used for determining the phenotypes, a thorough analysis of these genomic regions would be critical in understanding their roles in influencing antimicrobial susceptibility.

A second plasmid we have found to be associated with imipenem susceptibility, *pAYP-A2*, contains multiple genes that were statistically associated with imipenem susceptibility by Wallace *et al.* [122], suggesting these genes may be in linkage disequilibrium, and highlighting a limitation of gene-centric analyses of antimicrobial resistance.

The study of AMR is critically dependent on collections of well-characterized isolates, and on the availability of high-quality genome assemblies for the isolates. Our analysis was enhanced by our use of scaffolding information generated from the paired-end sequencing data, as well as by our reconstruction of complete genome sequences for several isolates using a long read sequencing technology. Such resources enable the localization of genomic features associated with resistance within plasmids or specific chromosomal locations and allow for analyses of gene order/gene proximity that can reveal polygenic factors that impact resistance.

The strategy we have applied here does not rely on computationally-expensive multiple genome alignment approaches, rather focuses primarily on local similarities between genomes, and is, therefore, scalable and generalizable. We believe the approach we have used provides a template for similar analyses in other organisms, powered by the increasingly-available collections of isolates with well-characterized phenotypes. Such datasets will provide fruitful sources of hypotheses for future experimental work, and expand our understanding of the genomic factors that underlie resistance to antimicrobials.

2.4 Materials and Methods

2.4.1 Primary dataset

The *A. baumannii* isolates were obtained from a collection of patient surveillance samples collected as a part of a cohort study at the University of Maryland Medical Center (UMMC). These isolates were from patients treated in UMMC medical and surgical intensive care units. A total of 349 isolates were selected for evaluation. These isolates were cultured on selective media for *Acinetobacter* species and tested for susceptibility to imipenem using the Kirby-Bauer disk diffusion method. The phenotypes of these isolates were determined using the Clinical and Laboratory Standards Institute (CLSI) standard antimicrobial susceptibility testing guidelines: isolates with their susceptibility of at most 18 mm and at least 22 mm were considered imipenem-resistant and imipenem-susceptible respectively. The carbapenem susceptibility scores for these isolates, along with their NCBI identifiers, are provided in Supplementary Table 1 from [21]. Additionally, a reference strain (Genbank ID: NC_017162) was used for the genomic analysis.

2.4.2 Genome assembly, MLST analysis, gene prediction, and gene clustering

All sequence read archive (SRA) runs for the 349 isolates were downloaded (449 SRA runs) and *de novo* assembled with `SPAdes` (v3.13.0) (default settings) [144]. These genomes were further filtered to retain the high confidence contigs that had at least 10× sequence coverage. Accounting for the genome completeness in comparison with the reference genome, 426 assemblies were identified for the 349 *A. baumannii* isolates. Each of these 349 isolates was then represented by only one of its assembled genomes; the sequencing run corresponding to the representative assembled genome is the first of its associated runs mentioned in Table S1. The quality of these selected 349 `SPAdes` assemblies as well as those available from NCBI were measured using `QUAST` [145]—`SPAdes` assemblies were observed to be better in terms of number of contigs, misassembled contigs, unassembled contigs, variance betwween N50 and NG50 values, etc. (Supplementary Table 9 from [21]), and were chosen for subsequent analyses.

In silico multilocus sequence type (MLST) profiles were assigned to the isolates using the sequences for the MLST markers from the Pasteur typing scheme (*gltA*, *recA*, *cpn60*, *fusA*, *pyrG*, *rpoB*, and *rplB*) [137]. The genomes were scanned against the pubMLST database (`http://pubmlst.org/abaumannii`) using `mlst` [146].

Genes were then predicted from these assemblies using Prokka (v1.12) [147].

The gene annotations available for the *Acinetobacter calcoaceticus-Acinetobacter baumannii* (ACB) complex were downloaded from NCBI. These genes were clustered using `CD-HIT` (v4.8.1) with a length difference cut-off of 0.8 (-s 0.8) to create a custom protein database [148], which was then used for the gene prediction. The predicted genes were used to estimate the pan-genome for these isolates using `Roary` (v1.007002) with a minimum protein sequence similarity of 85% (-i 85) and gene paralogs were clustered together (-s) [49]. The pan-genome was also used for recognizing the assemblies that shared sufficient similarity for the downstream analysis. The known AMR genes were then identified by comparing the corresponding assemblies of these 349 isolates against `CARD` [58, 143, 149].

2.4.3 Selection of additional 13 isolates for chosen microbial clade analysis

Among the isolates that did not host the ten strongly correlated known AMR genes, two imipenem-susceptible isolates were genomically similar to 13 imipenem-resistant isolates within NCBI's databases. One of these imipenem-susceptible isolates belonged to a SNP cluster on the NCBI Pathogen Detection Isolates browser (Biosample: SAMN02384922, cluster-id: PDS000005681). Using this SNP cluster, 13 isolates were identified which had imipenem susceptibility information in the database. All 13 isolates were imipenem-susceptible and collected in clinical setting by the Multidrug-Resistant Organism Repository and Surveillance Network,

Walter Reed Army Institute of Research (MRSN, WRAIR). The sequencing reads for these 13 isolates were downloaded and *de novo* assembled as before. Similarly, the genes were predicted from these genomes using Prokka. The pan-genome was estimated using Roary for the combined set of 28 isolates—13 imipenem-resistant and 15 imipenem-susceptible.

For further assessment, the 13 isolates were acquired from WRAIR. The following subsections (2.4.3.1–2.4.3.3) describe the procedures undertaken on these 13 isolates for reculturing, phenotyping for carbapenem resistance, long read sequencing, and complete genome assembly.

2.4.3.1 Biosafety, media, culture, and DNA isolation

All work associated with the preparation and extraction of materials from the multi-drug resistant isolates were performed in a BSL-2 laboratory under the class II biosafety cabinet with appropriate PPE (disposable laboratory coat, gloves, N-95 respirators, and face shields) based on risk assessment. The *Acinetobacter baumannii* strains were grown using BD Bacto™ Tryptic Soy Broth (Soybean-Casein Digest Medium). Single colonies were inoculated in TSB and grown overnight at 37°C. Cells were centrifuged and the pellet was washing using saline 0.85%. RNA-free genomic DNA was isolated using DNeasy Blood & Tissue Kits (Qiagen, Inc), following the manufacturer's instructions.

2.4.3.2 Carbapenem Phenotype Testing

Antimicrobial susceptibility testing (AST) was performed for 12 of the 13 isolates obtained from WRAIR by broth microdilution and Etest (bioMérieux). For both broth microdilution and E-test (E-strips), we performed two culture passages for individual colony isolation on blood agar plates. A 0.5 McFarland suspension (10^8 CFU/ml) was prepared to make broth microdilution plates cultures and a culture lawn on blood agar plates for E-strips. AST broth microdilution panels were incubated in aerobic incubation (36 degrees C) based on CLSI protocols. We utilized the Center for Veterinary Medicine's AST meropenem and imipenem panels (CMV4AGNF and CMV2DW) respectively. The GN panel (CMV4AGNF) requires 4 QC isolates and the ESBL panel (CMV2DW) requires 5 QC isolates as quality control organisms. Vitek 2 Compact (V2C) (bioMérieux) was used for culture identification and BMD panels were read manually using the Vizion (Sensititre, Trek Diagnostics). Culture lawn on blood agar plates for E-strips were incubated upside down in aerobic incubation (36 degrees). Interpretive Criteria for categorizing susceptibility, breakpoints for both AST broth microdilution and E-strips were adopted from the CLSI document M100.

2.4.3.3 Long read sequencing and assembly

Six of the 13 *Acinetobacter baumannii* isolates, acquired from WRAIR, were sequenced on the Pacific Biosciences (PacBio) Sequel System (PacBio, Menlo Park, CA). Specifically, the DNA from the isolates were part of a 4-plex to construct multiplexed microbial SMRTbell library using the SMRTbell Template Prep Kit 1.0 (PacBio, Menlo Park, CA) according to the manufacturer's protocol. The isolates were ligated with a unique barcode using the SMRTbell Barcoded Adapter Complete Prep Kit-96 (PacBio, Menlo Park, CA). Afterwards, size selection was performed with BluePippin (Sage Science, Beverly, MA). The multiplexed and size selected SMRTbell library was then sequenced on the PacBio Sequel sequencer using PacBio Sequel V2.0 chemistry on one Sequel SMRT cell 1M v2 (PacBio, Menlo Park, CA), with a 10h movie collection time. Raw sequencing data was demultiplexed by running the Demultiplex Barcodes application with the symmetric mode in SMRTLink v.7.0.1 (PacBio, Menlo Park, CA). The adapter sequence was trimmed and filtered out during the demultiplexing process. De novo assembly for the isolates was done using the PacBio hierarchical genome assembly process `HGAP`4.0 [150]. The genomes of the 6 isolates were checked manually for even sequencing coverage, circularized by `Circlator` 1.5.5, and polished by "Resequencing" in SMRT Link v.7.0.1 to ensure $>$ 99.99% mean consensus concordance [151].

2.4.4 Comparisons of specific AMR associated genes from prior findings

Similar to this study, a previous study by Wallace *et al.* had reported their findings based on genomic analysis of some other *A. baumannii* isolates collected as a part of the UMMC surveillance program [122]. Their analysis observed eight gene (centroid) sequences to be unique to carbapenem resistant isolates; these genes have not been previously considered to be AMR genes. The corresponding gene sequences were located in our assembled genomes using `BLAST`. The aggregated presence-absence counts for these eight genes is shown in Table 2.3.

Additionally, some other gene sequences were marked to be unique to these phenotypes within isolates grouped by their location of isolation—perirectal and sputum; these genes have also not been previously characterized as AMR genes. These comprised a total set of 385 gene sequences. We considered these genes to be strongly correlated with the resistance phenotypes, if they were reported to be unique to the corresponding resistance phenotype both overall or by source of isolation. The presence of these genes was determined across the available isolates using `BLAST`, and is shown along with the Fisher's exact test statistics in Supplementary Tables 2 and 6 from [21].

2.4.5 Genomic structural variants detection and lineage estimation

The assembled contigs from `SPAdes` assemblies were scaffolded using `MetaCarvel` with `-keep True` parameter and `-bsize 10` (without repeat resolution) to get the orientations of the contigs with at least 10 mate pairs linking these contigs [152]. The relative orientations of the contigs, in conjunction with the coordinates of the genes, predicted using `Prokka`, facilitated the prediction of the potential presence of the insertion sequence IS*Aba1* (Accession: EF571004) in the upstream regions of various beta-lactamase genes.

The lineage of the isolates was estimated using `Garli` based on a single nucleotide polymorphism (SNP) matrix computed with the reference-based `CFSAN SNP Pipeline` (v2.1.0) using the reference genome (Genbank ID: NC_017162) [43, 153]. The `CFSAN SNP Pipeline` takes pair-end reads from each isolate along with the reference genome and provides a SNP distance matrix representing the pairwise proximity between the given isolates in terms of the SNP distance. This distance matrix was then used to hierarchically cluster the isolates with single linkage.

Structural variants within the aforementioned 28 similar genomes were detected via multiple whole genome alignment with `Mauve` (snapshot_2015-02-25 build 0) [38]. `Mauve` constructs Locally Collinear Blocks (LCBs) to detect regions that are conserved across all or a subset of the given genomes. In addition, a custom script (using `PRAWNS` [154]) was used to detect the presence of shared k-mers (exact

matching substring of length k) between these genomes and the genomic positions of these k-mers were used to locate the conserved contiguous regions. The figures for the gene cassette and the genomic locations of variants on the reference genome were made using the `DNA Features Viewer library` [155].

The presence of genomic regions across the available isolates was ascertained using `BLAST` (`blastn` v2.8.1) or `nucmer` program from the `MUMmer4` package [156, 157, 158]. The genomic region under consideration was deemed to be detected if the corresponding match resulted in the sequence identity of at least 90%.

2.4.6 Association with resistance to carbapenems

We combined the genes from the comprehensive antibiotic resistance database (CARD) and genes that have been reported to have associations with carbapenem resistance; this provided 2650 genes that we take as the known antimicrobial resistance genes [58]. These known AMR genes were evaluated to check if their presence was statistically significantly correlated with the resistant isolates. We used Fisher's exact test from the `scipy.stats` Python3 package to get the significance statistics for each gene [159]. The entries of the 2×2 input matrix for the Fisher's exact test corresponded to the number of isolates that were resistant and would have the predicted gene, those that are resistant and would not have the gene, and similar respective counts for the non-resistant isolates. A gene cluster was deemed to be significantly associated with resistance if the test resulted in a p-value ≤ 0.01. Since the

number of resistant isolates largely exceeded the susceptible ones, we additionally constrained that a gene cluster was deemed significantly associated with resistant isolates if the test resulted in odds ratio ≥ 5 and was detected in ≤ 10 susceptible isolates. Among the 28 genetically similar isolates discussed before, owing to the high degree of genomic similarities, we constrained that a genomic variant (SNP or the conserved region) is strongly associated with a resistance phenotype if it is missing in at most one isolate of that phenotype (say resistant) and present in at most one isolate of the other phenotype (say susceptible).

Chapter 3: Plasmid-mediated Multidrug-Resistant *Shigella*

This chapter contains material previously published in Genomic Drivers of Multidrug-Resistant Shigella Affecting Vulnerable Patient Populations in the United States and Abroad [20], which was a joint work with Jay Noboru Worley, Maria Hoffmann, Kristen Hysell, Amanda Garcia-Williams, Kaitlin Tagg, Sanjat Kanjilal, Errol Strain, Mihai Pop, Marc Allard, Louise Francois Watkins, Lynn Bry. JNW performed biological experiments, performed bioinformatic and statistical analyses, crafted figures, and wrote the manuscript. KJ performed the large-scale comparative genomics and most of the bioinformatics analyses using high-performance computing, with contributions from ES, and MP. MH and MA sequenced the bacteria and the plasmids. KH provided the case report. AGW, KT, and LFW provided epidemiological data and analysis. LFW and AGW provided public health information and messaging. SK provided case data and research guidance. LB wrote and edited the manuscript and provided direction.

3.1 Introduction

Sexually transmitted *Shigella sonnei* and *S. flexneri* have been reported among men who have sex with men (MSM) in the United States since 1974 [160, 161, 162]. HIV positive MSM are at increased risk for sporadic shigellosis, and infection with HIV is hypothesized to be associated with a range of factors that can contribute to transmission of shigellosis among MSM [163]. Multiple US outbreaks of shigellosis among MSM have occurred over the past decade, resulting in calls for active additional testing, surveillance and health education efforts focused on MSM, patients, and healthcare professionals to address this public health threat [164, 165, 166, 167].

Antimicrobial resistance (AMR) in *Shigella* isolates is frequently mediated by plasmid-borne genes in addition to chromosomal determinants [168]. Isolates of multi-drug resistant (MDR), defined as resistant to at least three classes of antibiotics [169], *Shigella* from MSM in New York City have increasingly shown decreased susceptibility to the macrolide azithromycin [170], and a large proportion of *Shigella* isolates with decreased susceptibility to azithromycin have been shown among HIV positive men [171]. These findings are concerning given azithromycin's role in empiric treatment of infectious diarrhea and use for *Shigella* resistant to fluoroquinolones, penicillins, and sulfamethoxazole/trimethoprim [172]. However, in contrast to these other antibiotic classes, macrolides are not included in most automated clinical microbiology tests for drug resistance in *Enterobacteriaceae*. The lack of CLSI-validated clinical breakpoints for interpretation of macrolide results in

Shigella further confounds timely interpretation and reporting of resistance [173].

Azithromycin-resistant *S. sonnei* carrying the *mphA* gene have been found in the US since 1987 [174]. Plasmid *pKSR100*, recently associated with *S. sonnei* and *S. flexneri* infections among MSM, carries *mphA* [175, 176, 177, 178]. Insertion sequence (IS) IS*26* has been identified as an important factor in the evolution of *mphA* within this plasmid family [178].

We identified an MSM patient with MDR *S. sonnei* that demonstrated resistance to multiple antibiotic classes, including macrolides [169]. Integrated genomic and epidemiologic analyses of 2097 US-based *Shigella* cases, with functional *in vitro* studies, identified strain and plasmid-specific drivers of drug resistance, risks for spread, and potential for further acquisition of resistance. The clear identification of strain- and plasmid-level drivers of resistance and spread directed targeted clinical actions to improve rapid diagnosis and to support public health efforts to further reduce spread.

3.2 Results

3.2.1 Clincal case

A 50-60 year old HIV-positive self-reported MSM male reported two weeks of watery, blood-streaked diarrhea, left-lower quadrant abdominal pain, and 25-pounds

of weight loss over a month, without fever or chills. He had been off antiretroviral therapy for four years, had traveled to Central America one year prior, and had recent same-sex sexual activity prior to the onset of symptoms. The patient's white blood cell count was 6,000/μL with 19% eosinophils and a CD4 cell count of <50/µL. His HIV viral load was >150,00 copies/mL, and he tested positive for CMV viremia and *Strongyloides* IgG. An abdominal CT scan showed proctocolitis.

S. sonnei strain, SBJ-9962 (BioSample: SAMN07663113, PacBio Assembly: CP053751–CP053763, PulseNet ID: MA_17EN5067), was isolated from stool cultures. The patient was initially treated with ciprofloxacin, but switched to IV ceftriaxone after susceptibilities showed resistance to ciprofloxacin (>32 μg/mL), ampicillin (>32 μg/mL), sulfamethoxazole/trimethoprim (> 320 μg/mL), with susceptibility to ceftriaxone (<1 μg/mL). Upon request, an azithromycin Etest showed a MIC of >256 μg/ml. Symptoms resolved after four days, and he was transitioned to oral cefpodoxime. Antiretroviral therapy was resumed.

The patient was discharged to home on cefpodoxime, but several days later developed recurrence of diarrhea and abdominal pain. Repeat stool cultures again isolated *S. sonnei* with the same susceptibilities, producing the isolate SBJ-9961 (SAMN07662568). He was continued on oral cefpodoxime, but symptoms persisted. At that time, he was restarted on 2g of ceftriaxone IV daily. Symptoms improved after one week of therapy and repeat stool cultures were negative. After two additional weeks of intravenous ceftriaxone stool cultures remained negative.

3.2.2 National infrastructure for pathogen genomic surveillance identifies diverse plasmid replicons in MSM *Shigella*

The strains SBJ-9961 and SBJ-9962 were submitted to the GenomeTrakr Network [179], and to CDC's PulseNet network [180], distributed networks of laboratories in the US and abroad that utilizes whole genome sequencing for pathogen detection and analysis. De-identified genomic data from isolates is deposited to NCBI from both sources for access by local hospital, public health, and governmental personnel to support surveillance and active outbreak investigations from local through international levels.

A high-resolution, closed genome of strain SBJ-9962 identified 12 distinct plasmids including pMHMC-012, which showed sequence similarity to plasmid spA from *S. sonnei* strain Ss046, a plasmid with widespread distribution among *S. sonnei* (Figures 3.1, 3.2) [181, 182]; and pMHMC-004, which showed high similarity to pKSR-100. Six plasmids, including the MDR plasmids pMHMC-004 and pMHMC-012, contained known origins of transfer (Figure 3.2). Both pMHMC-003 and pMHMC-004 encoded putatively intact type IV secretion systems. Genomic analysis from NCBI's pathogen isolates browser indicates that SBJ-9961 and SBJ-9962 are clonally related, and they have no known phenotypic differences.

Functional studies of the host range for the two MDR plasmids showed that pMHMC-012 could be transferred by conjugation to S. sonnei and *E. coli*, while

pMHMC-004 could only be transformed into several *S. sonnei*, including recently-isolated clinical strains, by electroporation, but neither *E. coli* DH5α nor *E. coli* J53 (Supplementary Dataset 2 from [20]).

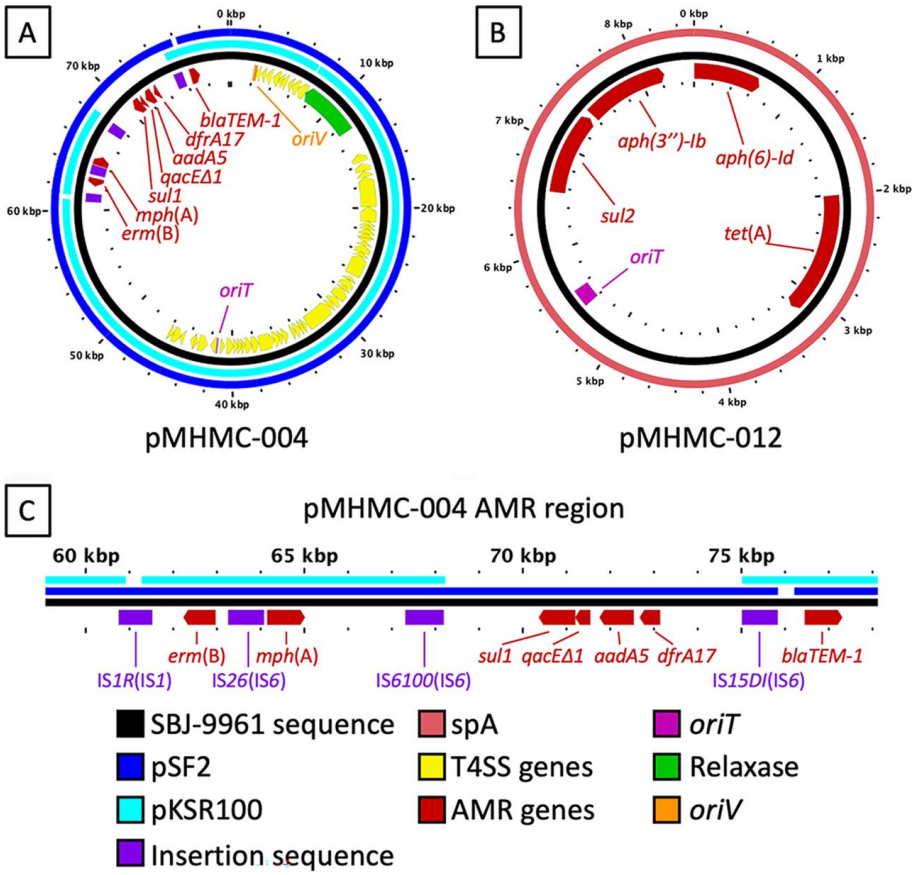


Figure 3.1: pMHMC-004 and pMHMC-012 are related to known multi-AMR gene plasmids. `BLAST` identity $\geq$ 80% with $\geq$ 100 bp length is shown by colored bars exterior to the plasmid backbone in black. Genes and genetic features are labeled as directional arrows and boxes, respectively, in the inner circle. (A) pMHMC-004. (B) pMHMC-012. (C) Details of the AMR gene region of pMHMC-004 with IS elements labeled, IS element family in parenthesis.

3.2.3 Antibiotic resistance profile

S. sonnei SBJ-9961 demonstrated resistance to seven antibiotic classes: aminoglycosides, fluoroquinolones, macrolides, penicillins, sulfonamides, and tetracyclines

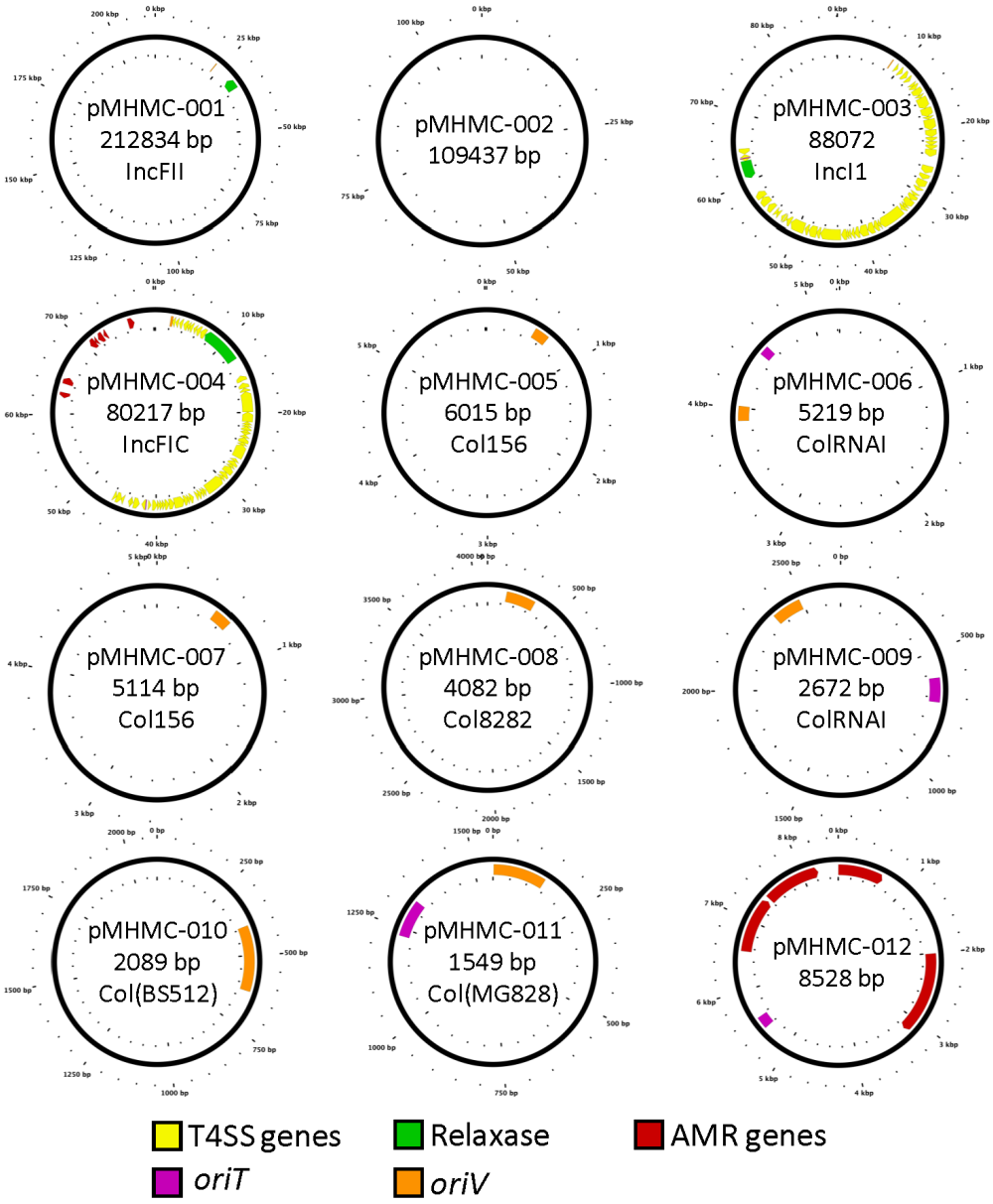


Figure 3.2: Twelve plasmids of *S. sonnei* SBJ-9962. Genes and genetic features are labeled as directional arrows and boxes, respectively. Only pMHMC-003 and pMHMC-004 encode a putative T4SS, while those plus pMHMC-006, pMHMC-009, pMHMC-011, and pMHMC-012 encode a known origin of transfer. Two plasmids—pMHMC-004 and pMHMC-012—carry known AMR genes. Most plasmids encoded a typical replicon: the larger plasmids had Inc type replicons and the smaller plasmids had Col type replicons.

(Supplementary Tables S1-S2 from [20]). Based on plasmid electroporation into the putatively pan-susceptible strain ATCC strain *S. sonnei* ATCC 25931, pMHMC-

004 conferred resistance to macrolides, penicillins, sulfonamides, and sulfamethoxazole/trimethoprim, while pMHMC-012 conferred resistance to tetracycline, streptomycin, and sulfonamides. Both plasmids alone confer an MDR phenotype, defined as resistant to antimicrobials in at least three antibiotic classes, but the strain remained susceptible to cephalosporins, colistin, beta-lactam combination agents, and several other classes of antimicrobials (Table S2 from [20]) [169].

3.2.4 Antimicrobial resistance determinants

SBJ-9962 carried chromosomal *dfrA1*, and *sat2* AMR genes which promote resistance to trimethoprim and streptothricin, respectively. It also carries the fluoroquinolone resistance mutations *gyrA* S83L and D87G, and *parC* S80I, mutations also identified in California outbreak strains in 2014 [174]. Two of the 12 plasmids carried known resistance genes: pMHMC-004 encoded *ermB*, *mphA*, *sul1*, *qacE*Δ*1*, *aadA5*, *dfrA17*, and $bla_{\text{TEM-1}}$; and pMHMC-012 encoded *aph(6)-Id*, *tetA*, *sul2*, and *aph(3")-Ib* (Figure 3.1). pMHMC-004 harbored four intact ISs, three of which belong to the IS*6* family, and one to the IS*1* family (Figure 3.1C). IS*26* and IS*6100*, both IS*6* family members, flank *mphA* (Figure 3.1C).

3.2.5 SBJ-9962 belongs to an *S. sonnei* clade that harbors high plasmid replicon counts

The genome sequences enabled improved analyses of *S. sonnei* outbreak strains in the US and abroad, particularly of SNP cluster PDS000033428. Members within this cluster averaged 5.6 unique replicons among their plasmids (Figure 3.3). However, the sub-clade within PDS000033428 that includes SBJ-9962 showed a much high number of replicons, including the 8 unique replicons found in SBJ-9962. SBJ-9962 contains an insertion in *cas6e*, a component of the CRISPR/CAS system which may play a role in its capacity to maintain a high plasmid count [183].

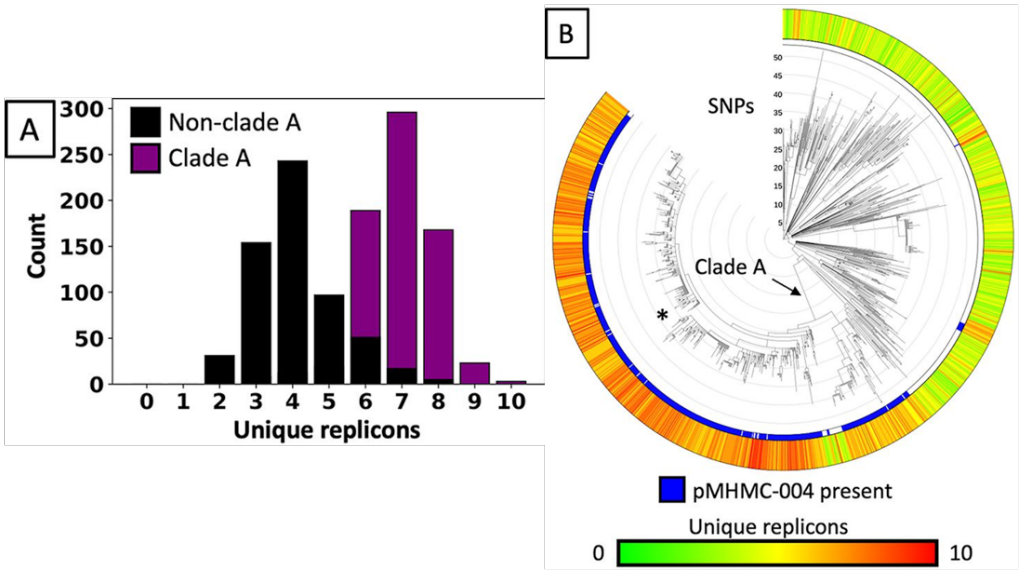


Figure 3.3: Unique replicon counts within PDS000033428. The *Shigella sonnei* SNP cluster PDS000033428.133 from the NCBI Pathogen Detection Isolates Browser, including isolate SBJ-9962 (*) with 8 different replicons, contains a subclade of strains with an elevated number of plasmid replicons.

3.2.6 Distribution of pMHMC-004 among *S. sonnei*, *S. flexneri*, and *E. coli*

Approximately 20% of both *S. flexneri* and *S. sonnei* isolates showed carriage of pMHMC-004 or pKSR100 homologs (Table 3.1). In contrast, only two of 42,465 E. coli isolates harbored any of these plasmids, one of which was an adherent-invasive *E. coli* (AIEC) LF82 (NC_011993.1), and the other was cultured from an asymptomatic case of bacteriuria (ABU) E. coli 83972 (NC_017631.1) [184, 185].

Species	**Neither**	**pMHMC-004**	**pKSR100**
S. flexneri	2785 (81%)	527 (15%)	117 (3%)
S. sonnei	4312 (81%)	873 (16%)	120 (2%)
E. coli	42463 (100%)	2 (0%)	0 (0%)

Table 3.1: pMHMC-004 carriage in strains of *Shigella* and *Escherichia coli*

Among *Shigella*, pMHMC-004 occurred in all *Shigella* clades (Figure 3.4), with multiple instances of plasmid acquisition and loss. Additional strains across clades carried gene regions with homology to portions of pMHMC-004 but at thresholds below those for calling pMHMC-004 or pKSR100 presence. Given the found sequence diversity for the pKSR100-related plasmids, heterogenous plasmid populations exist with broadly shared genetic cassettes [177, 178].

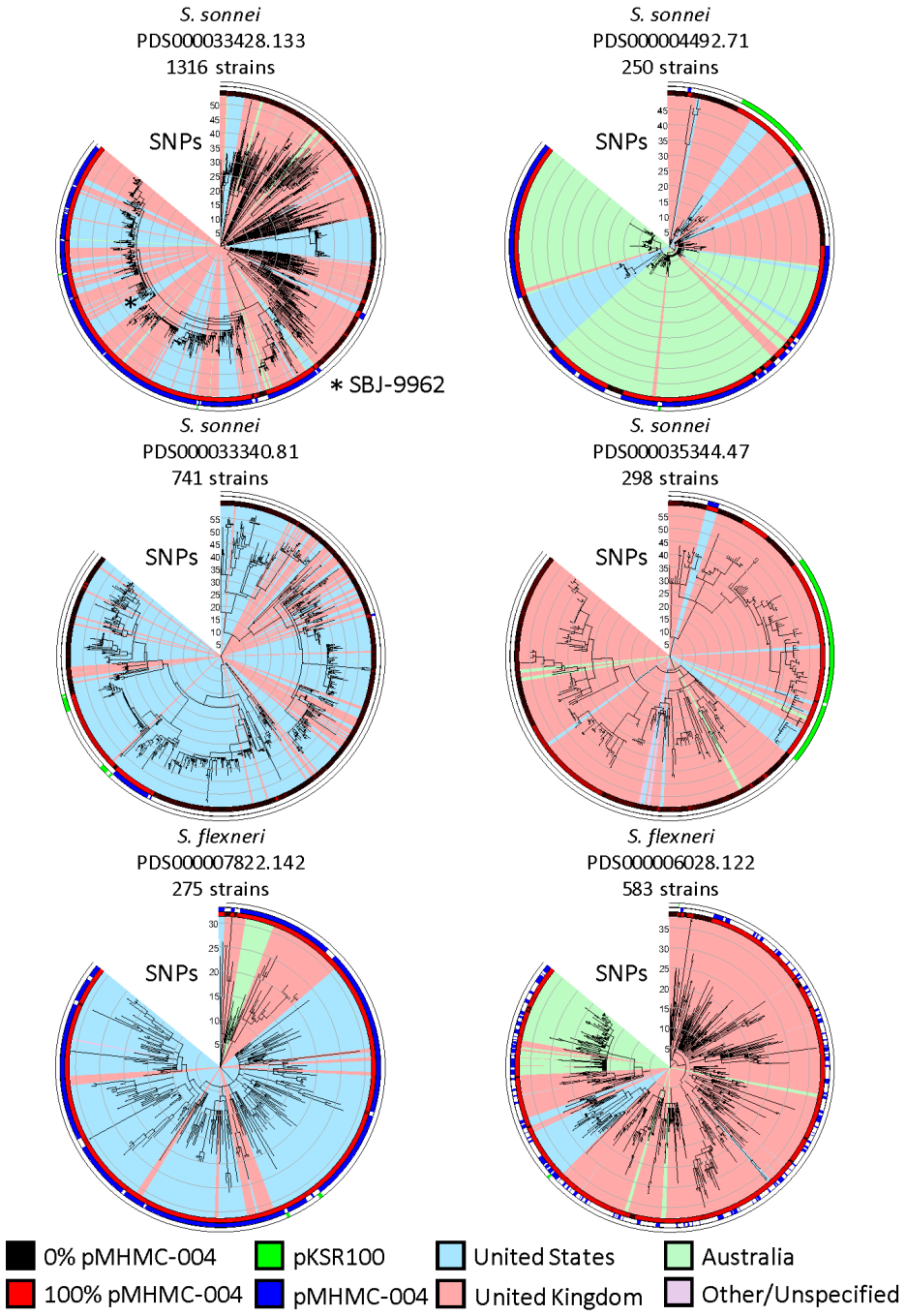


Figure 3.4: Large SNP-defined groups of *S. sonnei* and *S. flexneri* intercontinental transmission and plasmid acquisition and loss events. pMHMC 004 and pKSR100 presence is indicated in the exterior two rings. Total plasmid coverage is shown by a gradient in the first ring outside the dendrogram. Country of origin is as the United States, United Kingdom, Australia, and Other/Unspecified.

3.2.7 Intercontinental strain and plasmid transmission

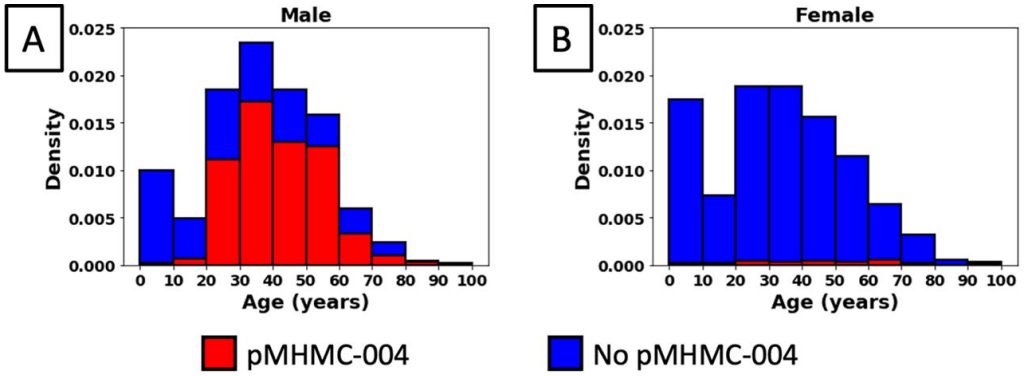
Shigella isolates among all SNP clusters defined by NCBI at the time of access originated primarily from the United Kingdom (2130 isolates), the United States (2097 isolates), and Australia (321 isolates), with 132 isolates from other countries or that were of unspecified origin (Figure 3.4). Isolates within each SNP cluster occurred across the three predominant countries, and subclades within these clusters showed strong geographical biases, suggesting localized spread with genetic divergence from a common ancestor. Cluster PDS000033428, which includes SBJ-9962, is particularly diverse in origin with repeated occurrence of the strain between the United States and the United Kingdom. These findings illustrate the occurrence of plasmid/strain complexes across continents, and demonstrate putative instances of intercontinental transmission.

3.2.8 pMHMC-004 and related plasmids have association with men in the United States

Epidemiologic analyses of 1883 US-based *Shigella* cases with known demographic data identified strong association of the MDR *S. sonnei* strain and plasmids with men (Figure 3.5). The average age of patients linked to isolates with pMHMC-004 was 41.3 ± 13.8, while the same for those linked to isolates without the plasmid averaged 33.0 ± 19.9 (p-value $= 9.9 \times 10^{-22}$). The discrepancy in age was largely

due to higher numbers of children (ages 0–18) in the latter category—only 11 of 308 (3.6%) isolates carried the plasmid. Similarly, of cases involving women, only 31 of 318 (9.7%) isolates carried pMHMC-004. However, among isolates from men, the 604 of 1257 (52%) cases carried the plasmid. In total, of all cases involving the plasmid, 604 out of 646 cases (93%) with both age and sex recorded involved men. The observed difference in this distribution was highly significant ($\chi^2 = 320$, dof = 2, p-value = $4.1e^{-70}$).

The largest SNP cluster, PDS000033428, which includes SBJ-9962, has five large clades, three of which include the majority of cases seen in men, and that also carry pMHMC-004 (Figure 3.6, Fig. S3, Supplementary dataset 3 from [20]). There is limited occurrence of this plasmid in clades not associated with infections in men.



C

	With pMHMC-004	No pMHMC-004	p-value
Age	41.3 ± 13.8	33.0 ± 19.9	$9.94e^{-22, t}$
Demographic			
Adult male (N, %)	604 (32.1%)	653 (34.7%)	
Adult female (N, %)	31 (1.7%)	287 (15.2%)	$4.07e^{-70, \chi}$
Children (N, %)	11 (0.6%)	297 (15.8%)	

$\chi = \chi^2$ test, t = t-test

Figure 3.5: Demographic distribution of *Shigella* isolates and pMHMC-004 plasmid presence. Normalized histogram showing the proportion of isolates with or without pMHMC-004 by age for (A) male patients, and (B) female patients. (C) Univariate analysis of plasmid carriage likelihood by age and demographic groups.

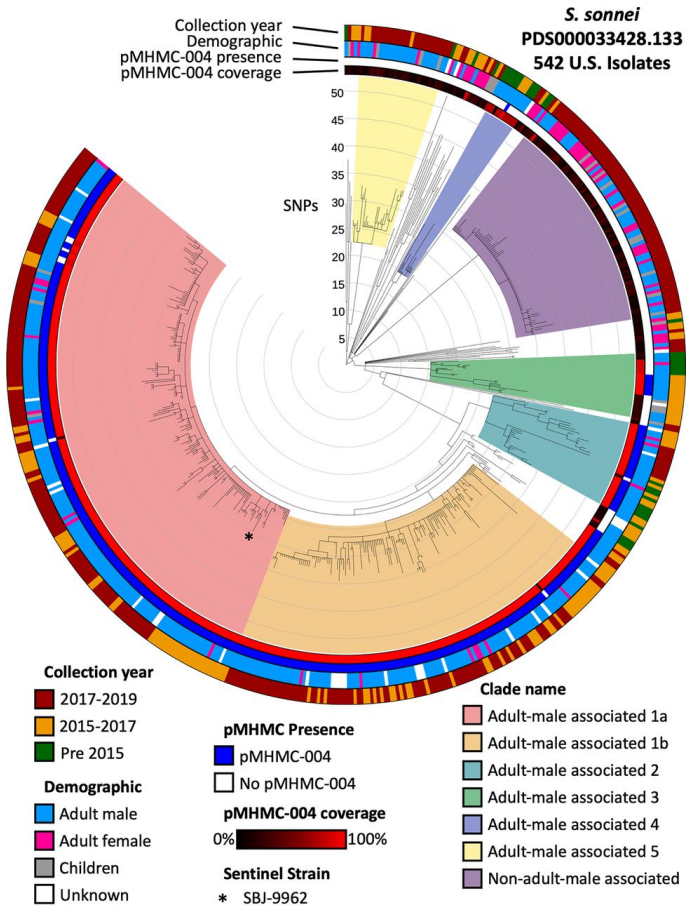


Figure 3.6: SNP cluster PDS000033428.133 includes isolates a recent and large outbreak of MDR *Shigella sonnei* in the US. The rings, from interior to exterior, represent pMHMC-004 coverage, pMHMC-004 presence, demographic group, and year of isolation. Clades of interest within this SNP cluster are color coded directly on the tree.

3.2.9 pMHMC-004 and pMHMC-012 show variable carriage of AMR genes over time

Significant variation occurred in the AMR genes seen in clades associated with outbreak cluster PDS000033428 (Figure 3.6, Fig. S3, Supplementary dataset 4 from [20]). After points of acquisition of macrolide resistance genes, sporadic isolates lacking *mphA* or *ermB* occurred, suggesting genetic excision events involving

the three IS*6* family ISs. Several strains lacked some or all genes bounded by these elements, indicative of multiple possible re-arrangements involving these IS*6* elements. The variability in MDR genes illustrates the degree of heterogeneity in the pKSR100 family of plasmids, attributes that are not unique to this or other *Shigella* plasmids, including spA/pMHMC-012.

3.3 Discussion

We present the first functional and genomic epidemiologic analyses of *Shigella* strains seen predominantly among men in the United States, suggesting transmission among MSM, and further define how these components fit more broadly within *Shigella* cases globally [186, 187]. *Shigella* spp. can quickly acquire and develop antibiotic resistance [188]. Notably, outbreak strains seen among men can carry many different plasmids, as highlighted by the 12 unique plasmid replicons identified in SBJ-9961. pMHMC-004 belongs to an emerging family of MDR plasmids that includes pKSR100, and has shown frequent acquisition and loss of antimicrobial-resistance genes commonly mediated by recombination events among homologous insertion sequences and other acquired repetitive elements [176, 178].

pMHMC-004 showed significant association in *Shigella* cases of adult men, including the MSM case that started our investigation, and not in other patient populations. This combined plasmid/pathogen complex significantly complicated our patient's treatment; chromosomal resistance to fluoroquinolones combined with

plasmid-mediated resistance to ampicillin, sulfamethoxazole/trimethoprim and azithromycin rendered the usual oral therapies ineffective and necessitated intravenous treatment. The receptivity of *Shigella* to plasmid acquisition raises further concerns for acquisition of additional resistance genes including extended-spectrum and other beta-lactamase enzymes that would further limit treatment options [189, 190, 191, 192].

While pMHMC-004 encodes a putative Type IV F-pilus, we were unable to conjugatively transfer the plasmid to susceptible recipients under laboratory conditions. However, pMHMC-012 was transferrable by conjugation, demonstrating the existence of a functioning conjugative F-pilus in this isolate. Electroporation experiments demonstrated a primary host range limited to other *Shigella* species and not to other *E. coli*, a finding further supported by genomic epidemiologic analyses that identified only two instances of pMHMC-004 homologs in >40,000 *E. coli* genomes, strengthening the idea that this plasmid has a limited host range (Table S2, [20]).

Relatives of the pMHMC-004 plasmid [177, 178] demonstrate variation in their antibiotic resistance gene profiles, suggesting that this family of plasmids provides an adaptable backbone for transmission of antibiotic resistance across *Shigella*. Clinical microbiologic phenotypic testing for antibiotic resistance in strains infecting vulnerable patient populations is paramount to insure proper treatment and, when appropriate during outbreaks, the implementation of measures for prevention. Ongoing national and international efforts for genomic surveillance furthermore provide means to actively monitor AMR gene acquisition to inform clinical, public health and food safety agencies for appropriate monitoring, diagnostic and therapeutic op-

tions to insure optimal and timely treatment. As part of surveillance activities, collection of demographic and behavioral data is essential to further characterize populations disproportionately impacted by MDR *Shigella*.

Our sentinel strain and others in the current US-based outbreak had putative macrolide resistance via *mphA* [176]. Given the current lack of CLSI clinical breakpoints to call macrolide resistance in *Shigella* species, labs may opt to proactively test *Shigella* with minimal inhibitory concentration (MIC)-based methods, to provide a minimum antibiotic concentration to direct clinical therapy. Longer-term, understanding the current drivers of antimicrobial resistance are best incorporated in ongoing national and international efforts to provide timely frameworks for diagnostic laboratories in support of patient care. These issues also highlight the need for reflex cultures if culture-independent diagnostic testing (CIDT) is used as a primary modality to diagnose of *Shigella* infections, particularly for patients who don't clear infection or who may be at higher risk for infection with MDR strains.

We note that drug-resistant cases of Shigellosis are not unique to MSM patient populations [164, 166, 176, 177]. Healthcare providers should consider the potential for drug-resistant *Shigella* in gastroenteritis patients who are immunocompromised and have not recovered, or in any case failing to respond to antimicrobial therapy. Education for patients with MDR shigellosis to prevent transmission through a range of modes, including food, water, and both sexual and non-sexual person to person contact, are important to further limit pathogen spread. Among MSM, where the disease appears to be more prevalent, physicians should take sexual histories into

account in providing prevention information. While multiple pathogens and clinical diseases can present with such symptoms, standard clinical microbiologic methods for identifying enteric pathogens, and for susceptibility testing of identified *Shigella*, are needed to provide appropriate diagnostic information to guide patient care and management.

3.4 Materials and methods

3.4.1 Clinical case

An MDR *S. sonnei* was identified in the Brigham and Women's Hospital Clinical Microbiology Laboratory and flagged by Infection Control for genomic analyses through the Partner's Pathogen Genomic Surveillance program (IRB protocol 2011-P002883; L. Bry) to identify drivers of antimicrobial resistance and relatedness with outbreak *Shigella* seen in the Northeast [193, 194]. Details about this case and the two isolates from it are presented in the Results.

3.4.2 Bacterial isolation and maintenance

Shigella isolates were isolated by stool culture on Hektoen enteric agar (Remel, Lenexa, KS) and speciated by VITEK 2 (Biomérieux, Durham, NC). Rifampicin-resistant mutants of clinical strains, used for functional studies, were created by

plating a 10 µl loopful of cells from LB agar (Becton, Dickinson and Company, Sparks, MD) onto LB agar with rifampicin overnight at 37 ºC. The following antibiotics were used for selections: ampicillin (MilliporeSigma, Saint Louis, MO), 200 μg/ml; rifampicin (G-Biosciences, Saint Louis, MO), 100 μg/ml; tetracycline (MilliporeSigma, Saint Louis, MO), 15 μg/ml. Bacterial strains used in this study are listed in Table S4 from [20].

3.4.3 Plasmid Transfer Analyses

Bacterial transformations were done as previously described [195], with additional screening done on CHROMagar MH Orientation agar (CHROMagar, Paris, France) and MacConkey Enteric agar (Remel, Lenexa, KS) to confirm species and strain phenotypes.

For conjugation and electroporation studies of plasmid transfer, donor and recipient strains were grown overnight on LB agar at 37 ºC, with selective antibiotics whenever appropriate. Cells were suspended in sterile LB to an OD600 of 0.1, then donor and recipient mixed 1:1, and 50 µl was plated onto LB agar. The reactions were incubated 16-20 h at 37 ºC, 30 ºC, or room temperature to evaluate temperature-dependence of conjugation. Reactions were resuspended into 5ml sterile LB using a sterile loop and diluted before plating. 30 μl of dilution was plated onto half a selective or non-selective agar plate for determining cell concentrations.

Electroporation was performed as described with the following modifications [196]. Cells were grown on solid media and resuspended in 300 mM sucrose (MilliporeSigma, Saint Louis, MO), and 2 mm cuvettes (Thermo Fisher Scientific, Waltham, MA) were used with a 2.5 kV, 25 mF, 200-Ω program and 100 μl of prepped cells. DNA for electroporation was purified using a QIAfilter Plasmid Midi Kit (Qiagen, Germantown, MD) with 100 ml of late log phase growth culture grown in dual selection with ampicillin and tetracycline at 37 ^{o}C. 1 μl of plasmid DNA solution (400 ng) was used in electroporation reactions.

3.4.4 Antibiotic susceptibility testing

Kirby-Bauer, Etests, and broth microdilution studies were performed within Clinical and Laboratory Standards Institute (CLSI) guidelines [173, 197, 198]. Azithromycin resistance levels were determined by Etest (bioMérieux, Durham, NC).

3.4.5 Genomic analyses

The Partners Pathogen Genomic Surveillance program is a node on the FDA's GenomeTrakr Network and submitted the strain for genomic analysis. DNA isolation and Illumina MiSeq sequencing were performed as described [193]. PacBio sequencing was performed as previously described [199]. Sequence accessions used in genomic analyses are listed in Supplementary Dataset 1 from [20].

Sequences were analyzed for resistance genes using the Bacterial Antimicrobial Resistance Reference Gene Database (PRJNA313047) and `BLAST` [200]. AMR genes were screened for using the thresholds of 80% protein sequence identity and 60% gene length, keeping the best hit per genome locus. All hits had >95% protein sequence identity and length. Replicons were identified using the PlasmidFinder reference sequences and with homologous plasmids called at 80% nucleotide identity and 60% coverage [201]. In the case of pMHMC-011, which had 92% sequence identity and 53% replicon coverage for Col(MG828), the replicon is interrupted by the beginning of the sequence. Plasmid-encoded type IV secretion system elements were identified using oriTfinder [202]. GenBank files with all features annotated were created using Biopython [203].

Sequencing reads from the *E. coli* and *Shigella* isolate genomes in the NCBI Pathogen Detection Isolates Browser were downloaded from the Sequence Reads Archive (SRA) in May 2019. The reads were *de novo* assembled using `SPAdes` with the `--careful --cov-cutoff auto` options to generate draft-level genome assemblies [204]. Putative plasmids were evaluated using `blastn` comparisons between the complete plasmid sequences and the assembled genomes. 80% sequence level identity was used as a cutoff for mapping reads to plasmids in calculating how coverage of reference plasmid sequence in draft genomes. Due to the nature of similarity between the plasmids pMHMC-004 and pKSR100, a minimum plasmid coverage of 95% was used. The dfrA17-¬sul1 resistance region was also used to differentiate pMHMC-004 and pKSR100 type plasmids (Figure 3.1C) [176]. `Kraken2` was used

to confirm strain species assignments [205].

For comparisons of pMHMC-004 to related plasmids pKSR100 and pSF2, GView was used on default settings of greater than 80% nucleotide sequence identity and an E-value of less than $1e^{-10}$ to determine overall sequence alignment and coverage of pMHMC-004 to the two reference plasmids [206]. In-tact ISs were detected using `ISfinder`'s `BLAST` function using 80% reference sequence coverage and 80% nucleotide identity as cutoffs for pMHMC-004 and pMHMC-012 [207].

3.4.6 Phylogenetic analyses

Phylogenetic relationships were obtained from the NCBI Pathogen Detection Isolates Browser. Visualizations with plasmid sequence alignment were made with the Interactive Tree of Life [208].

3.4.7 Epidemiological data management and statistics

US isolates had year of isolation, patient age, and patient sex information collected for analysis when available. Sexual orientation or sexual behavior data were not available for US isolates. Patients were classified into one of three demographic groups: men are patients who were of male sex and at least 18 years of age, women are patients who were of female sex and at least 18 years of age, and children were all patients under 18 years of age. The proportion of men was used as an

indicator to suggest MSM transmission. Previous work has used gender distributions to identify excess cases of shigellosis among men, which can suggest potential MSM transmission [186, 187]. For determination of the significance of patient age by isolate pMHMC-004 carriage, a t-test was used. A χ^2 test was used to evaluate the significance of demographic group between patient isolate pMHMC-004 carriage. T-test and χ^2 tests were performed using `SciPy` [159].

Chapter 4: Whole-genome population genomics using PRAWNS

As of writing this thesis, this chapter contains material under submission in ***PRAWNS****: Compact pan-genomic features for whole-genome population genomics [154], which was a joint work with Hugh Rand, Errol Strain and Mihai Pop. KJ, HR, and MP designed the study. KJ developed the entire code, performed all bioinformatic analyses, and created the figures. All authors analyzed the results and wrote the manuscript.*

4.1 Introduction

Rapid advances in sequencing technology have led to the availability of genomes for many bacteria, viruses, and fungi. The study of these genomes has been crucial in improving our understanding of different organisms; it is transforming public health and food safety monitoring, and has led to the development of sequence-based infectious disease programs [13, 14]. Methods developed for genomic analysis of small numbers of bacterial genomes, such as the alignment, gene-based, and k-mer approaches described in more detail below, are not able to keep pace with the

analysis demands of an ever increasing number of isolates produced through public health surveillance.

Genomic comparisons provide valuable insights into the evolutionary biology of organisms and allow a characterization of variation by, e.g., horizontal gene transfer (HGT) events, insertions-deletions (indels), and translocations. Such genomic variations can cause important phenotypic variations, including antimicrobial resistance (AMR) and pathogenic variation [11]. Analysis of genomic variations benefit from a pan-genome representation—genomic features present across a set of genomes are aggregated into a single unified representation [15] which readily permits assessing the genomic features potentially influencing phenotypic variation [12]. Large-scale genome comparisons (also referred to as population genomics or comparative genomics) permit the association of such genomic variants with the phenotypic variation due to geography, environment, AMR, and as found in case-control studies [20]. The understanding of relevant genomic factors and mechanisms underlying phenotypic variation enhances our ability to monitor phenotypes in important applications such as microbial food safety, disease invasiveness, and AMR [24].

Comparison of multiple closely-related genomes can be performed in a number of ways. The most obvious one is via whole genome alignment, and well-known tools for this include `Mauve` [38], `Mugsy` [39], and `Cactus` [40]. Another approach is gene-focused, and uses gene prediction followed by gene clustering to estimate the core and accessory genes. Prominent tools include `PanSeq` [47], `PGAP` [48], and `Roary` [49]. Other faster approaches rely on exact matches to construct a colored

de Bruijn graph: these tools, such as `SplitMEM` [55], `TwoPaCo` [56], locate exact matching substrings of fixed length—called k-mers—from all genomes and aggregated them into a compacted graphical representation for all genomes. A recently developed multiple whole-genome alignment pipeline, `SibeliaZ` [209], uses `TwoPaCo` to construct a de Bruijn graph, followed by locally collinear blocks (LCBs) extraction with `SibeliaZ-LCB`, and finally running `spoa` for multiple sequence alignment [210]. Among the above approaches, the compacted de Bruijn graph based approaches generally scale well to large numbers of genomes without limiting themselves to variants only present in genes.

Recently, some other approaches, called bacterial genome-wide association studies (GWAS), have been developed specifically for genotype-phenotype correlation studies; prominent ones include `SEER` [65], `DBGWAS` [66]. These approaches employ an alignment-free comparison strategy: they extract k-mers from the genomes or sequenced reads and model the k-mer distributions—this can then be used to predict a genome phenotype based on its k-mer profile. These newer k-mer-based approaches are quite fast but at the expense of losing the genomic context.

All approaches discussed in the last two paragraphs have their strengths and weaknesses. The methods based on whole genome alignment do not scale adequately with current sequencing capacity (>100 genomes). Reference-guided approaches, (e.g. [43]), suffer from the inherent bias in the reference chosen as well as the need for a closely related reference genome; species with an 'open' pan-genome are a poor fit for a reference-based approach. Gene-based approaches handle the entire pan

genome, but suffer from gene prediction biases and gene clustering thresholds [211]; additionally, poor assembly quality of the genomes is problematic for the gene prediction process. Compacted de Bruijn graph approaches are fast, scalable, and not limited to genes. However, the graph output comprises a large number of exact matches ('unitigs') which makes the downstream analysis difficult. Finally, the alignment-free approaches have to rely on various heuristics to constraint the k-mers to be evaluated, in addition to losing the genomic context. None of the current approaches provide a full solution to the problem of providing a well-scaling algorithm that provides genomic context and supports genome-wide association studies easily.

Variation in genomic context of specific genetic features can drive phenotypic variation. However, such variation has been largely understudied due to the inability of existing genome comparison approaches to support an efficient comparative assessment of the genomic context. As such, analyses of genomic context have largely been limited to the relationship between genes and the changes in their promoter regions.

Promoter region changes can influence the expression levels of corresponding genes—an important antimicrobial resistance (AMR) mechanism. Certain AMR genes are upregulated due to the presence of insertion sequences, a type of mobile element, in their promoter regions [71]). Mobile elements are known to pose challenges to genome assembly algorithms and often get assembled in separate contigs in draft assemblies [21, 70]. The fragmentation in draft assemblies prohibits a comprehensive comparison of multiple genomes while accounting for such collocated

occurrences of genomic factors which may be influential on the resultant phenotype. Overall, for a thorough comparative analysis of several related genomes, we need an approach that is scalable, accounts for the variations at a whole-genome level, and provides a reasonably small number of genomic features which could be examined for a downstream assessment.

Here, we propose a novel approach, `PRAWNS`: a fast and scalable tool that generates an efficient representation of closely related whole genomes to provide a concise list of genomic features or sequence entities shared by a user-specified fraction of the genomes. `PRAWNS` can be parallelized over multiple threads and uses disk-based storage, enabling it to scale to thousands of genomes. `PRAWNS` relies on two main algorithmic innovations. First, we identify segments of DNA that are shared across the genomes being analyzed; we call these segments the *'conserved regions'*. The detection of conserved regions begins with the identification of exact matching regions which are later merged into *'metablocks'*. These metablocks may contain inexact matches, and reduce the number of features to be considered by an order of magnitude over individual exact-matching blocks. `PRAWNS` achieves efficiency by using exact matches and by organizing them into metablocks without the need of an actual alignment. Our second algorithmic contribution is a new type of genomic feature called *'paired regions'*—these are pairs of conserved regions collocated in multiple genomes but that may vary in distance between each other within different genomes. Using such paired features, scientists can assess the influence of the collocation of genomic segments on phenotypes.

4.2 Methods

4.2.1 Overview, notation, and definitions

PRAWNS takes as input a set of whole genomes $\mathbf{G} = \{G_1, \ldots, G_{|\mathbf{G}|}\}$, where $|\mathbf{G}|$ denotes the total number of input genomes. Optionally, it also supports contig orientations obtained from using scaffolding [152]. For each genome, we find the k-mers that are unique in it; these k-mers are then merged to get shared exact matching regions, referred to as *'blocks'*. Note that the blocks are similar to the the maximal unique matches (MUMs) used in many other genome alignment approaches [36], except that they need not be maximal in each genome that contains them. Next, we determine the 'neighborhood' for each block—a block is defined to be in the neighborhood of another block if the separation between these blocks is at most a threshold δ nucleotides in some genome G_i; such a pair of blocks is called a 'candidate neighboring pair'. We employ these candidate neighboring pairs to identify the clusters of collocated blocks, referred to as the *'components'* of collocated blocks shared across multiple genomes (Algorithm 1). The collocated blocks from each component are merged to generate *'metablocks'* (Algorithm 2): each metablock is a chain of blocks that are present in the same order in at least a user-defined fraction (ϵ) of genomes under study. Additionally, we provide a set of *'retained blocks'* comprising the blocks which weren't merged into a metablock but are at least γ nucleotides long. The metablocks and retained blocks together constitute the primary

set of PRAWNS' features, and are referred to as the '*conserved regions*'—denoting the genomic regions shared between multiple genomes under study. We avail the conserved regions to compute an additional set of contextual genomic features—the paired regions. A paired region denotes a pair of conserved regions that are closely adjacent ($\leq \Delta$) with a consistent relative orientation in $\geq \epsilon \times |\mathbf{G}|$ genomes. All aforementioned thresholds are user-defined and their default values are provided in the subsequent sections.

We use a single coordinate system per genome: genomic coordinates start at 1, and increase from left to right. Contigs within an assembled genome are concatenated while maintaining the contig boundaries. For example, if the first contig is 1000 bp long, the next contig would start at 1001. A genomic feature f, i.e. a block or a metablock, located in genome G_i is represented by the tuple $\langle f.start_i, f.end_i, f.orientation_i \rangle$: $start_i$ and end_i are the left-most and right-most coordinates of f in G_i, while $orientation_i$ is 1 or 0 if f is located on the forward or reverse strand, respectively, in G_i. A $start_i = 0$ indicates that f is missing in G_i. Two features (metablocks or retained blocks) f_a and f_b constitute a paired region pr if they are closely adjacent ($\leq \Delta$) with a consistent relative orientation in $\geq \epsilon \times |\mathbf{G}|$ genomes. If pr exists in a genome G_i, then the value for pr in G_i denotes the nucleotide distance separation between f_a and f_b in G_i. If pr does not exist in G_i, then the corresponding value is 0.

4.2.2 Conserved regions detection

4.2.2.1 Exact matching blocks

We begin with detecting k-mers (k is user-defined, default: 25) from the given genomes—the idea is to find maximal exact matching regions shared across multiple genomes. We only consider unique k-mers from each genome, i.e. k-mers present only once within that genome, and construct a colored de Bruijn graph (cdBg) representation: a vertex denotes a unique k-mer, an edge exists in the graph if the corresponding $(k+1)$-mer formed using the adjoining two k-mers exists in the input genomes, and the edge-color denotes the genome membership for that $(k+1)$-mer. The cdBg is pruned to remove the vertices and edges that are present in fewer than $\epsilon \times |\mathbf{G}|$ genomes, where ϵ (default value: 0.05) can be user-specified. The pruned graph is 'compacted': a path P is compacted into a single vertex if all its vertices barring the first vertex have an in-degree 1, all but the last vertex have an out-degree 1, and each edge $e \in P$ has an identical set of edge-colors. Thus, each compacted vertex represents the corresponding exact matching region present in $\geq \epsilon \times |\mathbf{G}|$ genomes—the exact matching regions are referred to as *'blocks'*, denoted by $\mathbf{B}$. The blocks present in genome G_i are indicated by B_i.

Existing cdBg-based pan-genome tools typically stop at graph compaction and output the vertices and edges. Even with a small number, say 50, of a bacterial species' genomes, the vertex count usually exceeds 10^5. Downstream processing of

such a large number of features leads to computational bottleneck and statistical issues due to multiple testing. However, we observe the blocks to be often collocated in multiple genomes and could be represented as a single aggregated feature. We perform this aggregation over two phases: (i) identifying *'components'* of collocated blocks (ii) locating *'metablocks'* from these components.

4.2.2.2 *Components* of collocated blocks

Several blocks are often located within a proximity of one another across multiple genomes. We formulate the detection of such collocated blocks via construction of a K-nearest neighbors (KNN) graph and identifying the connected components in this graph (Fig. 4.1, Algorithm 1). Here, we make use of the fact that the genomes are linear (or circular): for each block, we can determine the blocks located within its 'neighborhood' on each genome. A block v is defined to be a candidate neighbor of another block u if the separation between these blocks is at most a threshold δ (user-defined, default: 5) nucleotides in some genome G_i; such a pair of blocks (u, v) is called a 'candidate neighbor pair'. For a candidate neighbor pair (u, v), $|G_{u,v}|$ denotes the number of genomes in which this candidate neighbor pair exists and is referred to as its genome occurrence count. For a block u, if $|G_{u,v}|$ has the highest genome occurrence count among all candidate neighbor pairs with u, then another candidate neighbor t is deemed a frequently encountered candidate neighbor of u if $|G_{u,t}| \geq (1 - \varphi) \times |G_{u,v}|$, where φ (default: 0.05) can be user-specified (see Section

1 of Supplementary Material). E.g. in Fig. 4.1, $|\mathbf{G}| = 4$ and $\varphi = 0.5$: for blocks b_i and b_j, b_j would be a frequently encountered candidate neighbor of b_i, if b_j is a candidate neighbor of b_i in at least $((1 - 0.5) \times 4 =)$ 2 genomes. Using these candidate neighbor pairs, we create a KNN graph: the vertices represent the blocks and each block has an edge directed towards at most K blocks corresponding to the blocks that formed the K-most frequently encountered candidate neighbor pairs. K is chosen to be a small integer and dependent on δ ($K \approx \sqrt{\delta}$); this ensures that the graph is sparse but collocated blocks remain connected. Ties are broken arbitrarily in cases with more than K neighboring blocks. For an edge $e(u, v)$, its edge weight $w(u, v) = |\mathbf{G}| - |\mathbf{G}_{u,v}| + 1$, i.e. the most frequently encountered candidate neighbors have the lowest edge weights. The KNN graph is pruned to retain the reciprocal nearest neighbors: u and v are reciprocal nearest neighbors if both u and v are K-nearest neighbors of each other, i.e. both the edges $e(u, v)$ and $e(v, u)$ exist in the KNN graph. Connected components are then extracted from the resultant graph $\boldsymbol{UG}$ and are referred to as the *'components'* of collocated blocks, denoted by $\mathbf{C}$. The run-time complexity of Algorithm 1 is $\mathcal{O}(|\mathbf{G}||\mathbf{B}| \log(|\mathbf{B}|))$.

4.2.2.3 *Metablocks*

'Metablocks' are chains of blocks that are present in the same order in $\geq \epsilon \times |\mathbf{G}|$ genomes. Algorithm 2 describes the process to identify the metablocks from the components of collocated blocks. First, we determine the genomes which

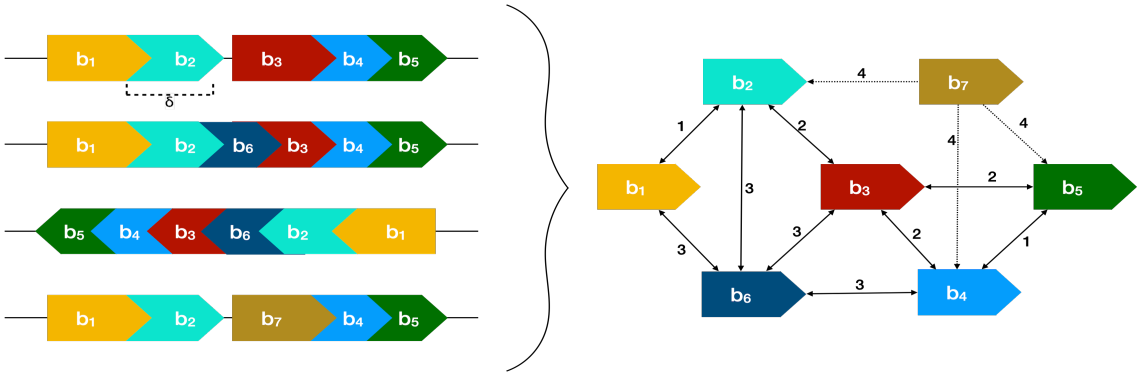


Figure 4.1: K-nearest neighbors graph ***DG*** (right) constructed from the blocks identified from four genomes (left) using the Algorithm 1. $K = 4$ and $\varphi = 0.5$. Blocks collocated $\leq \delta$ nt apart in some genome/s represent the candidate neighbors. The weight of an edge $e(u, v)$ is given by $w(u, v) = |\mathbf{G}| - |\mathbf{G}_{u,v}| + 1$; $|\mathbf{G}_{u,v}|$ signifies the number of genomes where v is a candidate neighbor of u. E.g. Blocks b_2 and b_4 are candidate neighbors in 2 of the 4 genomes, so $w(b_2, b_4)$ =4-2+1 =3. The dotted edges (unidirected neighborhood) are discarded to get ***UG***.

are likely to encode the metablocks by ascertaining the genome membership of their corresponding components: for each component c, we compute the number of blocks from c that are present within each genome. A component is considered to be present in a genome G_i if at least ω_c of its blocks (by default, 90%) are present in G_i. The corresponding subset of genomes containing the component c is denoted as $\widetilde{G_c}$. Next, to ensure a consistent ordering of blocks within a metablock region, we identify the *'core block pairs'* for each component: blocks u and v form a core block pair of component c, if u and v are adjoining pairs of blocks with a consistent relative orientation in each genome $G_i \in \widetilde{G_c}$. The core block pairs are a subset of candidate neighboring blocks. By design, adjacent core block pairs can have at most one block in common; core block pairs that share a block are extended to form *'chains'*. Chains from a component c are merged and extended, if they are at most a certain distance, μ (user-defined, default: 25), apart and have a consistent relative orientation in each

Algorithm 1: Algorithm for determining components of collocated blocks

Input: blocks $\mathbf{B}$, integers δ, K, $|\mathbf{G}|$, and fractions ϵ, φ

Output: a set of components of collocated blocks $\mathbf{C}$

$\mathscr{C} \leftarrow \emptyset$ ▷ Set of all candidate neighbor pairs

for $1 \geq i \geq |\mathbf{G}|$ **do**

  $CN_i \leftarrow \emptyset$ ▷ Candidate neighbor pairs from genome G_i

  let $b_j, b_l \in B_i$

  if $b_l.start_i – b_j.end_i \leq \delta$ and $\left|\left\{b_m \middle| b_j.start_i < b_m.start_i < b_l.start_i\right\}\right| < K, \forall b_m \in B_i$ **then**

    $CN_i \leftarrow CN_i \bigcup \{(b_j, b_l)\}$ ▷ New candidate neighbor pair

  Repeat from (4) until no new candidate neighbor pair found

  $\mathscr{C} \leftarrow \mathscr{C} \bigcup CN_i$

$\mathbf{N} \leftarrow \emptyset$ ▷ Set of candidate neighbor pairs sorted by decreasing genome occurrences

for all candidate neighbor pairs $(b_j, b_l) \in \mathscr{C}$ **do**

  add (b_j, b_l) to $\mathbf{N}$

  $|\mathbf{G}_{b_j,b_l}| \leftarrow |\mathbf{G}_{b_j,b_l}| + 1$ ▷ update genome occurrence count

Construct the K-nearest neighbors graph $\boldsymbol{DG} = (V, E)$ using $\mathbf{N}$ ▷ Vertices V correspond to distinct blocks and the weighted edges E denote the neighborhood relationship between the blocks

for all candidate neighbor pairs $(b_j, b_l) \in \mathbf{N}$ **do**

  let $g_{max}(b_k) \leftarrow \max |\mathbf{G}_{b_k,b^*}|$ ▷ Maximum genome occurrence count among all candidate neighbor pairs with b_k

  if $|\mathbf{G}_{b_j,b_l}| \geq \epsilon \times |\mathbf{G}|$ genomes **then**

    if $|\mathbf{G}_{b_j,b_l}| \geq (1 - \varphi) \times g_{max}(b_j)$ **then**

      Add directed edge $e(b_j, b_l)$ to $\boldsymbol{DG}$ with weight $w(b_j, b_l) = |\mathbf{G}| - |\mathbf{G}_{b_j,b_l}| + 1$

      Retain $\leq K$ out-edges from b_j with high edge weights

    if $|\mathbf{G}_{b_l,b_j}| \geq (1 - \varphi) \times g_{max}(b_l)$ **then**

      Add directed edge $e(b_l, b_j)$ to $\boldsymbol{DG}$ with weight $w(b_l, b_j) = |\mathbf{G}| - |\mathbf{G}_{b_l,b_j}| + 1$

      Retain $\leq K$ out-edges from b_l with high edge weights

$\boldsymbol{UG} \leftarrow ReciprocalNearestNeighborGraph(\boldsymbol{DG})$ ▷ Vertices u, v are connected if both u and v are K-nearest neighbors of each other. Edge weight is same as that in $\boldsymbol{DG}$

$\mathbf{C} \leftarrow$ connected components from $\boldsymbol{UG}$ ▷ Collocated blocks

return $\mathbf{C}$

genome $G_i \in \widetilde{G_c}$—the resultant merged region constitutes a *'metablock'* (see Section 1 of Supplementary Material). A larger μ can yield longer metablocks with several mismatches corresponding to the substitutions and indels between adjacent chains. If contig orientations (via scaffolding) are provided, chains can also be merged to generate composite metablocks that span across contig boundaries. Algorithm 2 has a run-time complexity of $\mathcal{O}(|\mathbf{G}||\mathbf{B}|)$.

Algorithm 2: Algorithm for constructing metablocks from components of collocated blocks.

Input: blocks $\mathbf{B}$, components $\mathbf{C}$, contig orientations $\mathbf{O}$ (via scaffolding, optional), integers μ, $|\mathbf{G}|$, and fractions ϵ, ω_c
Output: a set of metablocks $\mathbf{M}$

$\mathbf{M} \leftarrow \emptyset$ ▷ Set of metablocks
for each component $c \in \mathbf{C}$ **do**
  $\widetilde{G_c} \leftarrow$ genomes where c is deemed present
  $cp_c \leftarrow$ core block pairs for c
  $chains_c \leftarrow$ Merge and extend adjacent core block pairs in cp_c
  $M_c \leftarrow \emptyset$ ▷ *Metablocks* from component c
  for each chain $\mathscr{c}$ in $chains_c$ **do**
    if relative orientation of $\mathscr{c}$ and previous metablock is consistent and their separation is $\leq \mu$ in each genome $G_i \in \widetilde{G_c}$ **then** ▷ Uses $\mathbf{O}$ if available
      EXTEND_AND_UPDATE(previous metablock)
    else
      CREATE(new metablock) ▷ Create new metablock corresponding to $\mathscr{c}$ and insert in M_c
  $\mathbf{M} \leftarrow \mathbf{M} \bigcup M_c$
return $\mathbf{M}$

Empirically, we observe that many components comprise just one or two blocks, potentially originating from spurious k-mers (see Section 3 of Supplementary Material). We, therefore, confine the metablock detection to the components with ≥ 3 blocks to get longer metablocks and remove noise. Observe that each

block can be a part of only one component and can, hence, contribute to at most one metablock. However, some blocks may not be merged into a metablock either because they formed components with at most just one more block or the block was present but the corresponding component wasn't deemed present in certain genomes. Unmerged blocks, longer than a threshold γ (user-defined, default: 50), are referred to as 'retained blocks'; the metablocks and the retained blocks together comprise the *'conserved regions'*—the primary set of `PRAWNS`' pan-genome features. The conserved regions output constitute their FASTA sequences and two tabulated files denoting the genomic coordinates and binary presence/absence in respective genomes.

4.2.3 Paired regions detection

Identical genomic regions can give rise to different phenotypes depending on their orientation and association with other genomic regions [71, 212]. To capture such paired interactions, we developed the *'paired regions'* feature in `PRAWNS`: each feature corresponds to two conserved regions, r_1 and r_2, and their relative orientations, such that r_1 and r_2 are separated by at most a small distance Δ (user-defined, default: 50) nucleotides in $\geq \epsilon \times |\mathbf{G}|$ genomes. Note that for a pair of conserved regions r_1 and r_2, a feature is constructed for each distinct pair of relative orientations. Once the paired regions are identified, they are checked for their presence (same orientation) in the remaining genomes but farther apart ($> \Delta$ separation).

The output for the paired regions is represented by two comma-separated files: one file provides its binary presence/absence in each input genome while the other contains the nucleotide separation between the genomic coordinates of the constituent conserved regions. A negative value for separation corresponds to overlap between the conserved regions in the respective genomes. Observe that the output file containing the separations also allows assessing whether the distance between conserved regions is influential in the downstream assessments.

4.2.4 Implementation

`PRAWNS` (https://github.com/KiranJavkar/PRAWNS.git) is an open source code available under the GPLv3 license [154]. It is implemented in C++ and Python3 and works on Unix-like operating systems. The default parameters are calibrated for bacterial genomes. Detailed documentation for `PRAWNS` is available along with its source code [154].

`PRAWNS` is designed to work using limited main memory (RAM usage) with the availability of disk usage and supports parallelization. Each module of the tool—identification of blocks, components of collocated blocks (Algorithm 1), metablocks (Algorithm 2), paired regions—can be executed in parallel over tc cores (user-defined, default: 8) that access a shared disk space. If the maximum length of given genomes is N and the total sequence length is n, then the run-time complexity to detect the pan-genome features is $\mathcal{O}(n \log N + |\mathbf{G}||\mathbf{B}| \log(|\mathbf{B}|))$. For closely

related genomes with roughly similar genome lengths ($n \approx |\mathbf{G}| \times N$), the run-time is linear in the number of input genomes.

4.2.5 Analysis

4.2.5.1 Genome Assembly

Sequenced reads for all genomes analysed were downloaded from NCBI and assembled with `SPAdes` (v3.13.0, default settings) [144]. In order to ensure better confidence in the genomic variants detected, the assembled genomes were filtered to retain only those contigs that had at least 10× sequence coverage.

4.2.5.2 Statistical significance testing

It is important to note that the statistical analysis described here is simply meant to demonstrate the use of the features generated by `PRAWNS`, and is not an integral part of `PRAWNS` itself. The exploration of different statistical approaches for conducting association studies based on `PRAWNS` features is beyond the scope of this manuscript. For genotype-phenotype correlation analysis, the statistical significance of the features was assessed using Fisher's exact test and the Benjamini-Hochberg procedure for false discovery rate (FDR) correction ($\alpha = 0.01$).

4.3 Results

4.3.1 Datasets

We benchmark the performance of `PRAWNS` on two bacterial whole genome datasets: (i) 362 *Acinetobacter baumannii* genomes (approx. 4 Mbp long) (ii) 664 *Streptococcus pyogenes* (Group A Streptococcus) genomes (approx. 2 Mbp long). Both datasets contain two different clinically important phenotypes. The *A. baumannii* genomes vary substantially in imipenem resistance [21]. The *S. pyogenes* genomes (obtained from population diversity studies) vary in their invasiveness as categorized by their isolation source [65] To assess the scalability, `PRAWNS` was executed on two additional datasets: (i) 4000 *Salmonella* Infantis genomes (bacteria, approx. 5 Mbp per genome) (ii) 107 *Aspergillus flavus* genomes (fungi, approx. 38 Mbp per genome) (Supplementary Table 1).

4.3.2 Methods compared

As of the time of writing this manuscript, only `SibeliaZ-LCB` [209] is able to generate a pan-genome representation at a whole-genome scale, can handle a large number of genomes, and processes a compacted de Bruijn graph to produce a reduced set of genomic features—the locally collinear blocks (LCBs). As the LCBs defined by `SibeliaZ-LCB` are similar to the conserved regions from `PRAWNS`, we

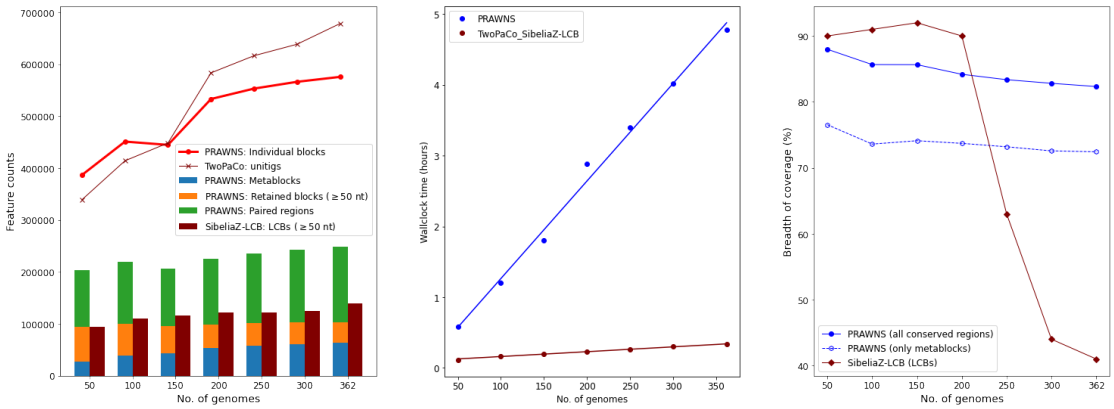
compared the performance of PRAWNS against SibeliaZ-LCB; we limit the multiple whole-genome alignment pipeline SibeliaZ to only the first two steps, i.e. compacted de Bruijn graph construction using TwoPaCo followed by LCB detection with SibeliaZ-LCB. However, it is important to note that SibeliaZ-LCB is intended for detection of homologous sequences that have an evolutionary distance to the most recent common ancestor (MRCA) of at most 0.09 substitutions per site, and does not account for potential rearrangements or larger structural changes.

We used the combination of TwoPaCo (v0.9.3) and SibeliaZ-LCB (v1.2.2) (hereafter referred to as TwoPaCo+SibeliaZ-LCB) for comparative benchmarking of PRAWNS' performance. TwoPaCo was run with filtersize 10. SibeliaZ-LCB was run with the parameters -m 50 -b 25 to maintain consistency with the default values used for PRAWNS parameter ϕ, γ, Δ. DBGWAS was run with default parameters and $k = 25$ [66]. All tools were run on a 64 core Xeon E5-2680 server running at 2.70 GHz and a total of 256 GB of RAM.

4.3.3 Performance

4.3.3.1 Scalability

Figure 4.2 compares the performance and feature counts on the *A. baumannii* dataset. The dataset (362 isolates) was randomly sampled without replacement to create smaller datasets of 50, 100, 150, 200, 250, and 300 genomes. The tools



(a) Pan-genome feature counts (b) Run-time (c) Breadth of coverage

Figure 4.2: Scalability performance using *A. baumannii* dataset. (a) Feature counts from `PRAWNS` and `SibeliaZ-LCB`. For comparison, the number of blocks (`PRAWNS`) and unitigs (`TwoPaCo`) are shown. The total number of conserved regions from `PRAWNS` is the combination of metablocks and the retained blocks ($\gamma = 50$). (b) Run-time performance of `PRAWNS` and `SibeliaZ-LCB`. (c) Median breadth of coverage by the conserved regions and LCBs.

were run with k-mer length 25 on 8 cores and the maximum memory usage was limited to 36GB. Figure 4.2a shows line plots for the counts of exact matching regions detected in the genomes: blocks from `PRAWNS` (dense red line) and unitigs from `TwoPaCo` (thin red line with $\times$'s). The bar plots represent the feature counts generated by `PRAWNS` and `SibeliaZ-LCB`. The merger of blocks into metablocks (blue bars) results in an order of magnitude reduction in their counts. The total counts of conserved regions, i.e. metablocks and retained blocks (orange bars), are comparable to that of the LCBs from `SibeliaZ-LCB` (maroon bars). Observe that the total counts of `PRAWNS`' features (conserved regions and paired regions) are still much smaller than that from mere de Bruijn graph compaction. Figure 4.2b shows that both—`PRAWNS` and `TwoPaCo+SibeliaZ-LCB`—can operate on large number of closely related whole genomes with a run-time linear in the number of

input genomes, but TwoPaCo+SibeliaZ-LCB is much faster than PRAWNS. The actual RAM usage for PRAWNS was under 8.3 GB while that for TwoPaCo+SibeliaZ-LCB was under 3GB for the complete set of 362 genomes. Figure 4.2c demonstrates the breadth of coverages for the genomes obtained using conserved regions from PRAWNS and LCBs from SibeliaZ-LCB. Here, we see a stark difference between the features obtained from the two approaches. The median breadth of coverage with PRAWNS' conserved regions is 88% for 50 *A. baumannii* genomes and slightly reduces with added genomes, reaching 82% for all 362 genomes. The LCBs from SibeliaZ-LCB account for a median breadth of coverage of 90% for 50 genomes; however, on larger datasets, the breadth of coverage drops rapidly and reached 41% using all 362 genomes.

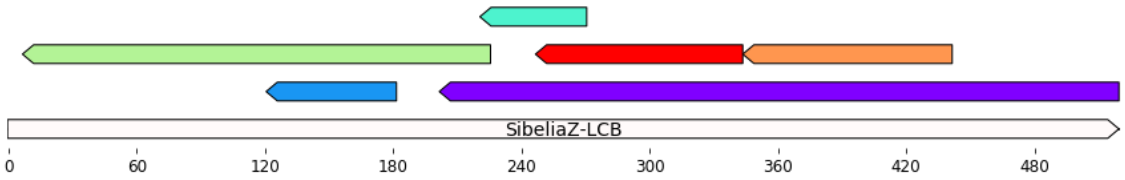
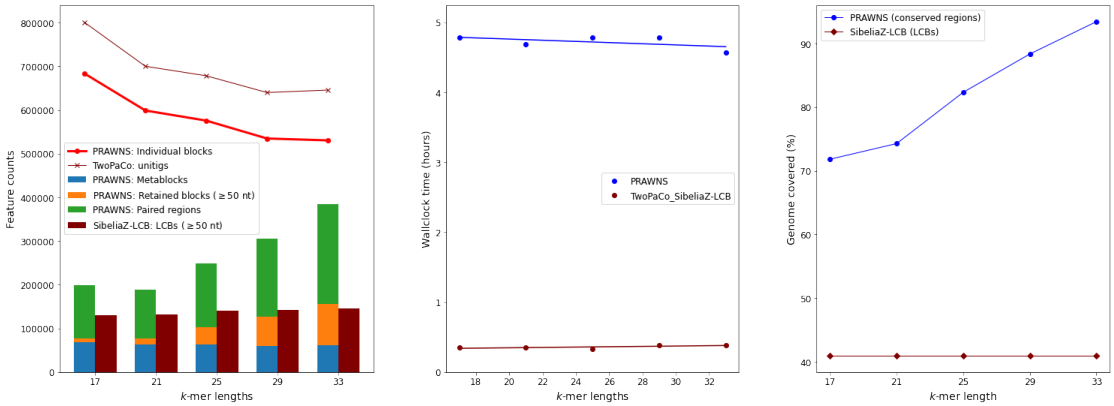


Figure 4.3: Conserved regions from PRAWNS aligned to a 520 nt LCB from SibeliaZ-LCB. The two longer conserved regions are metablocks while the remaining four are additionally retained blocks ($\gamma = 50$). The unique color assigned to each conserved region signifies distinct *presence vectors*, i.e. different genome memberships. SibeliaZ-LCB presumes this to be a contiguous homologous region and be deemed present or absent in the genomes.

Next, we explore the similarities and differences between the conserved regions and LCBs using the output for all 362 genomes. The LCBs were mapped to conserved regions using BLAST. 97,612 out of 102,623 (95.12%) conserved regions mapped to 113,527 out of 139,968 (81.11%) LCBs at above 90% identity. However, multiple conserved regions mapped to a single LCB and simultaneously also mapped

to multiple LCBs—suggesting redundancy in the genomic regions marked by different LCBs. Figure 4.3 shows one such example where `SibeliaZ-LCB` identified a single LCB of 520 nucleotide length, whereas `PRAWNS` decomposed this region into six conserved regions; the two longer ones are metablocks. The colors of these conserved regions denote the *presence vectors* (or equivalence classes or color classes) for these regions. We observe that these regions have different genome memberships and, therefore, should not be represented as a single collinear homologous region.



(a) Pan-genome feature counts (b) Run-time (c) Breadth of coverage

Figure 4.4: Impact of k-mer length using the *A. baumannii* dataset (362 genomes). (a) Feature counts from `PRAWNS` and `SibeliaZ-LCB`. The number of blocks and unitigs are shown for comparison. The total number of conserved regions from `PRAWNS` is the combination of metablocks and the retained blocks ($\gamma = 50$). (b) Run-time performance of `PRAWNS` and `SibeliaZ-LCB`. (c) Median breadth of coverage by the conserved regions and LCBs.

To demonstrate `PRAWNS`' scalability to thousands of genomes, we ran it on 4000 *Salmonella enterica* subsp. *enterica* serovar Infantis (*S.* Infantis)—all isolates were selected from an NCBI Pathogens SNP cluster (cluster ID: PDS000089910, Supplementary Table 1) and had a mean total genome length of 4.97 Mbp. `PRAWNS` was executed using default parameters on 16 cores each of 2.30 GHz Xeon E5-2650

processor with 50 GB maximum RAM limit. The execution completed in 24 hours (24.09 GB peak memory usage) and generated a pan-genome comprising 55,366 conserved regions (6,553 metablocks and 48,813 retained blocks), yielding a median breadth of coverage of 89.41%, and 86,175 paired regions.

To assess the scalability of `PRAWNS` on longer (eukaryotic) genomes, we ran it on a fungal dataset comprising 107 *Aspergillus flavus* genomes downloaded from NCBI (see Section 2 of Supplementary Material). *A. flavus* contains 8 chromosomes and the genomes are approximately 38 Mbp each. Using the default k-mer length (25) on five cores each of 2.70 GHz Intel Xeon E5-2680 processor with 50GB maximum RAM usage limit, `PRAWNS` required 20 hours (40.26 GB peak memory usage) to generate the pan-genome comprising 858,081 conserved regions (427,698 metablocks and 430,383 retained blocks ($\gamma = 50$)) and 1,020,418 paired regions ($\Delta = 50$). As in the case of *A. baumannii* genomes, the run-time was linearly proportional to the number of *A. flavus* genomes.

4.3.3.2 Robustness to the choice of k-mer length

Figure 4.4 shows the influence of choice of k-mer length on the entire *A. baumannii* dataset. Figure 4.4a shows a steady decrease in the number of blocks with an increased choice of k-mer length. However, the number of metablocks remains largely the same irrespective of the k-mer length. The k-mer length, however, varies the counts for the longer retained blocks and paired regions: smaller k-mer length

yields fewer longer blocks and paired regions whereas larger k-mer length results in more longer blocks and paired regions. The LCB counts from `SibeliaZ-LCB` remain fairly unchanged by the k-mer length choice. As observed in Figure 4.4b, the run-time performances of both, `PRAWNS` and `TwoPaCo+SibeliaZ-LCB`, are robust to the choice of k. The actual RAM usage was consistent with before for all these configurations. Figure 4.4c shows the impact of k-mer lengths on the median breadth of coverage of the genomes by resultant conserved regions and LCBs. The increase in k-mer lengths contributes to higher breadth of coverage at the expense of more conserved regions that would need to be analyzed. The breadth of coverage by metablocks alone increased slightly from 70% to 76% with an increase in k-mer length from 17 to 33. The overall increase in the breadth of coverage can be attributed to the detection of more unique k-mers and longer blocks (and, hence, an increased count for the retained blocks) with an increase in k. In contrast, the breadth of coverage using LCBs from `SibeliaZ-LCB` remains at around 41%—unperturbed by k-mer length choice.

4.3.4 Applications

4.3.4.1 Genotype-phenotype correlation studies

`PRAWNS` generates a pan-genome representation which can identify the genomic features for downstream population genomic analyses. As a proof of principle, we

demonstrate the utility of PRAWNS for population genomic studies using genotype-phenotype correlation analyses.

Antimicrobial resistance in *Acinetobacter baumannii*:

Imipenem (a type of carbapenem antibiotic) resistance phenotypes were determined for 362 *A. baumannii* genomes using the Clinical and Laboratory Standards Institute (CLSI) antimicrobial susceptibility testing guidelines [21, 213]. Out of the 362 genomes, 305 were imipenem-resistant while the remaining 57 were imipenem-susceptible (Supplementary Table 1). The genomic features constructed with PRAWNS comprised 102,623 conserved regions (63,358 metablocks and 39,265 retained blocks) and 145,615 paired regions, while SibeliaZ-LCB generated 139,968 LCBs. In order to ensure consistent behavior with that of PRAWNS, rather than relying directly on the information provided by SibeliaZ-LCB, we aligned the LCBs to the genomes using nucmer [158] and deemed an LCB present in a genome if it aligned with $\geq 90\%$ identity and $\geq 90\%$ query coverage.

First, we focus on genomic features shared by almost all the genomes (core-genome features found in $\geq$99% genomes). The core-genome features from PRAWNS comprised 6641 conserved regions and 2208 paired regions. When mapped against the CARD AMR gene database [58] using BLAST, the conserved regions aligned to several resistance and virulence genes including $bla_{\text{OXA-51-like}}$, bla_{ADC}, *abeM*, *adeFGH*, and *adeIJK*. The core-genome features from SibeliaZ-LCB comprised 1035 LCBs; the resistance and virulence genes aligning to these regions included $bla_{\text{OXA-51-like}}$,

bla$_{\text{ADC}}$, and *adeIJK*.

Next, we examined the genomic features with statistically significant correlations with the imipenem-resistant phenotype (Fisher's exact test with Benjamini-Hochberg FDR correction). The significant features from `PRAWNS` comprised 39,665 conserved regions (33,941 metablocks and 5,724 retained blocks) and 75,241 paired regions, with a total sequence length of 3 Mbp, while those from `SibeliaZ-LCB` comprised 7,424 LCBs, with a total sequence length of 7.53 Mbp. The conserved regions with significant association with resistance aligned to several AMR genes, including all 10 AMR genes (*bla*$_{\text{OXA-23}}$, *msrE*, *mphE*, *ANT(3")-IIa*, *aacC1*, *aphA6*, *qacEdelta1*, *sul1*, *yafP*, and *xerD*) reported in our previous work [21] and known to be strongly correlated with imipenem resistance—demonstrating the utility of `PRAWNS`' features for such genotype-phenotype analyses. On the contrary, the significant LCBs aligned to only two of the 10 AMR genes strongly correlated with imipenem-resistance—*aphA6* and *xerD*—and missed many important AMR genes associated with imipenem-resistance.

In comparison with `DBGWAS` (q100) 4,068 significant conserved regions were identified as 'top' conserved regions (significant conserved regions with among the 100 lowest FDR-adjust p-values). These had a total sequence length of 388 kbp and aligned with 5 strongly correlated AMR genes: *bla*$_{\text{OXA-23}}$, *msrE*, *qacEdelta1*, *sul1*, and *xerD*. `DBGWAS` identified 19 subgraphs spanning a total sequence length of 175 kbp and also aligned to 5 strongly correlated AMR genes: *bla*$_{\text{OXA-23}}$, *msrE*, *mphE*, *ANT(3")-IIa*, and *xerD*. However, none of the significant nodes (unitigs) from these

subgraphs aligned with any AMR gene.

The significant features from PRAWNS also support the discovery of other important and potentially understudied genomic factors influencing the phenotype. When aligned with NCBI BLAST [157], the statistically significant conserved regions also mapped to several mobile genetic elements and gene promoter regions associated with antimicrobial resistance, including insertion sequences (IS*Aba1*, IS*Aba13*, IS*Aba17* families), phage related genes, and plasmids. Additionally, the conserved regions mapped to other genes, such as *vgrG* [214] and *tviB* [215], which are relatively understudied with regard to antimicrobial resistance. The paired regions also provide a new dimension to the analysis: 75,241 paired regions were identified to be statistically significant, suggesting the presence of genomic factors whose co-presence is significant.

A similar genotype-phenotype analysis is not possible with the LCBs from SibeliaZ-LCB (see Section 4 of Supplementary Material). Due to the redundancy in genomic regions associated with multiple LCBs, the downstream analysis based on LCB presence/absence does not identify the LCBs containing known AMR factors to have significant phenotypic correlations.

Invasiveness of *Streptococcus pyogenes*:

Next, we appraise the applicability of PRAWNS' features for other clinically important phenotypes. We ran PRAWNS on a dataset of 664 *S. pyogenes* (Group A

Streptococcus (GAS)) genomes obtained from population diversity studies [65]: the genomes were from two geographical locations—Fiji ($n=352$) and Kilifi ($n=312$)—and categorised as invasive ($n=174$, isolated from blood, cerebrospinal fluid (CSF) or bronchopulmonary aspirate) or non-invasive ($n=490$, isolated from throat, skin or urine) (Supplementary Table 1). There was a higher proportion of invasive isolates from the Kilifi collection (133/312, 43%) compared to those from Fiji (41/352, 12%).

The resultant *S. pyogenes* pan-genome consisted of 69,151 conserved regions (46,092 metablocks and 23,059 retained blocks) and 146,163 paired regions. The conserved regions aligned to the known *S. pyogenes* resistance genes: *lmrP*, *mefE*, *tetL*, and *tetM* [216]—these conserved regions were identified in the genomes of invasive as well as non-invasive isolates from both Kilifi and Fiji.

Next, we assessed the statistical significance of pan-genome features for their propensity to be in invasive isolates. Accordingly, 1666 conserved regions (966 metablocks and 700 retained blocks) were identified to be significantly correlated with the invasive isolates. These include the *tetM* gene encoded within a conjugative transposon (Tn*916*)—previously reported by [65]. Although this transposon was identified primarily within the Kilifi isolates (213/312), unlike findings by [65], `PRAWNS` could locate this transposon in six isolates from Fiji as well. Additionally, conserved regions with significant hits included those that aligned to the CRISPR *Cas9/csn1* and *csn2* genes, which are known to limit the prophage insertions in *S. pyogenes* and, in turn, contribute to invasive GAS infections [217]. Another significant hit corresponded to the *FtsK* gene—reported to be differentially expressed

in isolates that survive in human amniotic fluid (AF) [218]. Other significant hits included conserved regions that aligned to AMR and virulence genes, such as the lantibiotic modification enzyme *LanM* and tetracycline resistance gene *tetL* [219].

`PRAWNS`' pan-genome features enabled the detection of genomic regions associated with the phenotype of interest even in draft assemblies without any dependency on a curated database of genes or genomic variants. The significant hits facilitated the discovery of known genomic factors associated with phenotypes as well as locating putatively novel factors associated or contributing to the exhibited phenotype.

`PRAWNS`' features can also be used to train a machine learning model such as for genotype-phenotype classification. A machine learning model can also be trained using `PRAWNS`' features. We use the pan-genome features from *S. pyogenes* genomes to train a XGBoost classifier [220] and classify the isolates as invasive or non-invasive. The dataset ($n = 664$) was split into 80% training data and 20% testing data. Using just the conserved regions presence-absence as the features, the model predicted with a testing accuracy of 76% (precision: 76%, recall: 76%). Complementing the feature set with the presence-absence of paired regions improved the prediction accuracy slightly to 77% (precision: 77%, recall: 77%).

4.3.4.2 Population structure

One of the key requirements of population genomics is the ability to visualize or represent the population structure of the isolates collection. To demonstrate the applicability of `PRAWNS` for this requirement, we used the pan-genome features obtained from the *S. pyogenes* dataset. A weighted hamming distance matrix was computed using the presence-absence of the conserved regions; the weighted scaling was performed using the lengths of the differentially present conserved regions. This distance matrix can then be visualized as a newick tree (Figure 4.5).

The population structure tree highlights the similarities and differences between the isolates at a whole-genome scale (Figure 4.5). We observe a diverse mix in the isolates across the invasive phenotypes as well as isolation source. A subset of isolates from Kilifi and Fiji have high genomic similarities while some others form distinct clades comprising isolates from a single origin; some other clades place some invasive isolates in the proximity of some non-invasive isolates. With such a whole-genome-scale overview, we can handpick isolates for a targeted high-resolution downstream analysis.

4.4 Discussion

We developed `PRAWNS` to fill a gap in the current tool-kit available to scientists for assessing the association of bacterial genomic features with phenotypes, such

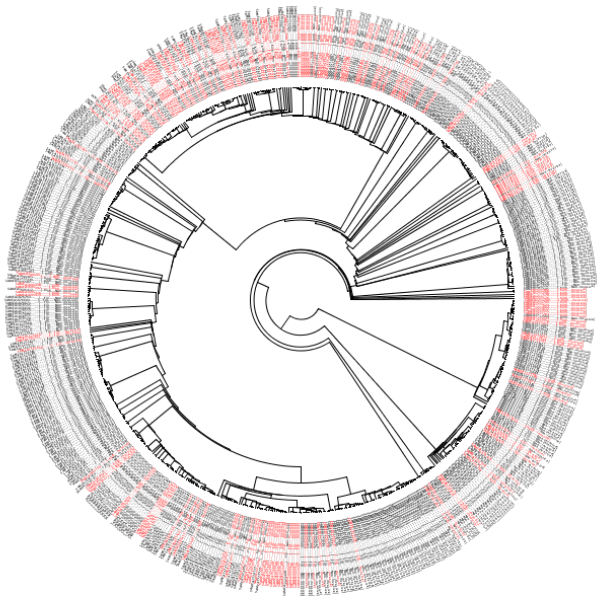


Figure 4.5: Population structure of 664 *Streptococcus pyogenes* isolates estimated using the conserved regions located using `PRAWNS`. The tree labels in red and black correspond to invasive and non-invasive isolates respectively.

as antibiotic resistance. Whole-genome alignment approaches do not scale beyond tens of hundreds of genomes, whereas the association studies relying on SNPs or k-mers require the assessment of features whose counts greatly exceed the number of genomes analyzed. Other studies that use genes as the features being associated with a phenotype may miss important factors such as mutations in promoter regions and interactions between adjacent genes. With `PRAWNS`, we present a middle ground in the form of scalability to large numbers of genomes (in the manuscript we show scalability up to 4000 genomes being aligned to each other) while maintaining a modest collection of features that facilitate the analysis of multi-chromosome isolates, allow for proper treatment of mobile genetic elements, and support the assessment of draft assemblies. Additionally, the provision to analyze the interactions between genomic regions can enable uncovering biologically important genomic factors, which have

been rarely characterized by multiple-genome association tools.

PRAWNS scales to a large number of genomes, and generates a concise set of features that are comparable in number with those produced by whole genome multiple alignment approaches, at a substantially lower computational cost. When applied to real biological data sets, PRAWNS' features can be effectively employed to gain insights into phenotype associations as we have seen in examining antibiotic resistance in *A. baumannii*, where we have recapitulated known resistance determinants and also revealed new potential associations [21].

Currently, PRAWNS does not explicitly consider SNPs and short insertion/deletion events; however, the metablock structure can be used as an anchor for constructing multiple alignments of genomic regions in order to identify such events, and we plan to implement such functionality in future versions of our code. Like many other tools, PRAWNS does not handle variation within repetitive regions of a genome. The ability to handle pairwise interactions between conserved genomic elements may provide an opportunity for us to define a unique context for each repeated genomic segment, thus potentially resolving this long-standing problem in multiple genome analysis. In its current implementation, PRAWNS relies on a straight-forward implementation of k-mers which allows ample opportunity for refinement in order to improve performance, e.g. using minimizers or other hash-based techniques, as well as ideas from compressed de Bruijn graphs.

Overall, PRAWNS provides an efficient and effective framework for exploring

genomic variations in large numbers of bacterial genomes, and we expect this tool to be a valuable part of the tool-kit used by scientists to analyze the rapidly increasing volumes of genomic data.

Chapter 5: Quasimetagenomic analysis of *Listeria monocytogenes* using long and short reads

This chapter contains material previously published in Evaluating the accuracy of Listeria monocytogenes assemblies from quasimetagenomic samples using long and short reads [99], which was a joint work with Seth Commichaux (co-first author), Padmini Ramachandran, Niranjan Nagarajan, Denis Bertrand, Yi Chen, Elizabeth Reed, Narjol Gonzalez-Escalona, Errol Strain, Hugh Rand, Mihai Pop and Andrea Ottesen. SC and KJ performed bioinformatic analyses, wrote the manuscript and created figures. AO and PR designed the experiment. AO, PR, YC and NGE performed microbiological and molecular lab work. ER created figures and organized data. NN and DB ran Opera-MS assemblies and analyses. SC, KJ, AO, PR, YC, NGE, ES, HR and MP edited the manuscript.

5.1 Background

5.1.1 State of the art for pathogen typing

Rapid response, whole-genome sequencing (WGS) networks such as GenomeTrakr [179], PulseNet [180], and the National Antimicrobial Resistance Monitoring System (NARMS) [221, 222] have revolutionized the strain typing and source attribution of bacterial pathogens and antimicrobial resistance (AMR) important to human and animal health. These programs have relied primarily on high throughput short-read sequencing data generated using the Illumina MiSeq platform. Accurate strain typing of bacterial pathogens using short reads is typically accomplished with SNP (single nucleotide polymorphism) and/or MLST (multi-locus sequence typing) analyses. Both can be performed directly on the raw reads or with assemblies of the raw reads. SNP analyses quantify the number of SNPs between a set of isolates and a reference genome [43]. High resolution MLST analyses involve identifying the profile of alleles for genes in the core genome and whole genome [223, 224], cgMLST and wgMLST, respectively. Both methods can differentiate between very closely related strains of *Salmonella enterica*, *Listeria monocytogenes*, *Escherichia coli*, *Staphylococcus aureus* and many other pathogens [225, 226, 227]. However, despite providing high resolution, SNP and cgMLST/wgMLST analyses do not analyze or require the entire genome assembly and, thus, miss aspects of genome architecture, such as the synteny of features and mobile elements with variable gene content [228].

5.1.2 The assembly of genomes using short and long reads

Ideally, complete genomes would be routinely sequenced and assembled *de novo* from outbreak samples for strain typing analyses. However, this is not yet possible in every situation. Although short reads can be sequenced with an error rate of less than 0.1% [229], these reads are typically 250 base pairs or less in length and cannot span many genomic repeat regions, resulting in fragmented assemblies that preclude the recovery of complete bacterial genomes [79]. In contrast, long read sequencing technologies like the Oxford Nanopore platform have higher sequencing error rates (∼13% [230, 231]), but can routinely produce reads that are over 10 Kbp, thus spanning genomic repeats and supporting the assembly of complete bacterial genomes and plasmids [232].

Although assemblies of nanopore long reads can generate genome-length contigs, they often have a large number of errors inherited from the reads. The hybrid assembly of Illumina short and nanopore long reads can remarkably improve the quality of the assemblies while maintaining syntenic contiguity [232]. A study of the assembly of several *Salmonella enterica* strains demonstrated that short read assembly followed by long read scaffolding, reconstructed genomes more accurately than using short reads or long reads alone [233]. Another study reconstructed entire genomes of Shiga-toxin producing *Escherichia coli* strains using nanopore long reads that were polished with Illumina short reads [234]; however, these assemblies had less accurate cgMLST typing compared to those using only MiSeq short reads,

despite the short read polishing.

5.1.3 Microbiological recovery of the target pathogen

Irrespective of sequencing technology, for applications such as the source tracking of bacterial pathogens, a fundamental challenge is the extraction of sufficient quantities of pathogen DNA to sequence in the first place. This is because pathogens frequently occur at low abundance in complex microbial communities, sometimes amongst large numbers of host cells, and/or in chemically challenging matrices. Current methods address this challenge by selective culture enrichment and pure colony isolation of the pathogens prior to sequencing and analysis. This approach, however, is labor-intensive and can take days to weeks to provide sufficient DNA for sequencing. While protocols and media formulations for the enrichment of *L. monocytogenes* vary only slightly between agencies (Food and Drug Administration (FDA), International Organization of Standardization (ISO), and the United States Department of Agriculture (USDA)), in-house FDA metagenomic and quasimetagenomic analyses of timepoints along recovery continuums from different starting matrices have demonstrated that enrichment dynamics and efficiencies vary according to chemical and microbiological features of the input matrix (i.e. different foods such as fresh produce, poultry, complex environmental samples, and varying initial loads (CFUs) of target pathogens) [235]. Community dynamics during all types of pathogen enrichments (e.g. *Salmonella enterica*, *Escherichia coli*, *Listeria spp.*)

are still poorly understood and co-enriching non-target species often compete with pathogens of clinical significance [236].

5.1.4 Metagenomics

Metagenomics is the direct sequencing of microbial communities [81] and, in theory, could replace culture enrichment for pathogen source tracking. Short read sequencing has been used extensively for metagenomics due to low error rates and high throughput, but cannot assemble many of the genomic and intergenomic repeats present in environmental DNA. In contrast, the long reads generated by nanopore sequencing platforms can resolve many of the genomic and intergenomic repeats. Recently, metagenomic studies have successfully used nanopore sequencing for rapid identification of dominant pathogens [237, 238] contributing complete assemblies for a small subset of the bacteria in the full metagenome [79, 239, 240]. However, achieving sufficient depth of coverage to assemble pathogen genomes directly from metagenomes is often prohibitively expensive.

5.1.5 Quasimetagenomics

A middle ground between the direct sequencing of samples and the sequencing of isolates from selective enrichments is quasimetagenomics, the sequencing of abbreviated recovery enrichments [79, 80]. Quasimetagenomics has been used by

FDA scientists since 2009 in efforts to recover pathogens from complex microbiomes such as outbreaks of *Salmonella* in tomatoes [78, 241], to better understand Latin cheese microbiota [242], to look at enrichments for Salmonella from cilantro [243], *E.coli* in flour [244], pathogens in seafood [245, 246, 247] and in the public health research response to the Blue Bell ice cream outbreak of 2015, which resulted in the dataset presented here [80, 236]. The first FDA ice cream work (2015) received a lot of attention in the food safety community and the quasimetagenomic approach was quickly emulated by other food safety research groups [75, 80, 248]. Many groups are moving the needle forward–demonstrating that strain level differentiation during an outbreak response can be achieved more rapidly with quasimetagenomic approaches [75, 248]. Here we build upon the first ice cream report [236] which demonstrated that a quasimetagenomic approach could recover the same quality of source tracking data much earlier than state of the art WGS approaches; and a second work which validated the bioinformatic SNP and cgMLST source tracking efficiency of the quasimetagenomic data [80]; and–presented here–the added value of GridIon long reads for circularization of genomes and plasmids.

5.1.6 Integrated microbiological, molecular and bioinformatic innovations that will move the field forward

Here, we provide a detailed benchmarking analysis for assessing how rapidly and accurately a targeted pathogen, *L. monocytogenes*, can be assembled from

quasimetagenomic samples using short and long read sequencing technologies. The evaluated assembly tools include those developed specifically for metagenomic assemblies (`MegaHit` [249] for short read assembly, `metaFlye` [250] for long read assembly, and `Opera-MS` [251] for hybrid assembly) as well as popular tools developed for long read genome assembly (`Canu` [252] and `Redbean` (formerly `wtdbg2`) [253]) and hybrid genome assembly (`HybridSpades` [254]). Additionally, we evaluated the impact of polishing with three tools: `Pilon` [255], `ntEdit` [256] (both were used to polish long read assemblies with short reads), and `Racon` [257] (was used to polish long read assemblies with long reads). The results of this study allowed us to point out the strengths and weaknesses in currently available tools and to make recommendations for future research.

5.2 Results

5.2.1 Characteristics of the sequencing data

The GridIon nanopore instrument generates sequencing data in batches of 4000 reads, denoted here as B_n for the n^{th} batch. The first 30 batches of GridIon reads, at each enrichment time, were used for this study, i.e., the first 120,000 reads corresponding to batches $B_1, B_2, \ldots, B_{30}$ (Figure 5.1). To analyze the quality of assemblies as a function of increased sequencing depth, each successive batch of reads was combined with the previous batches for assembly to form "cumulative

batches", denoted as $C_1, C_2, \ldots, C_{30}$, where $C_n = B_1 + B_2 + \ldots + B_n$ (Figure 5.1). To compare assembly results strictly based on sequencing technology, the number of base pairs for the MiSeq and GridIon data was normalized. Over a range of sequencing depths, MiSeq raw read files were partitioned into 30 corresponding batches of read pairs to match the cumulative batches by number of base pairs for GridIon reads. Table 5.1 records the total number of sequenced bases per C_{30} at each enrichment time.

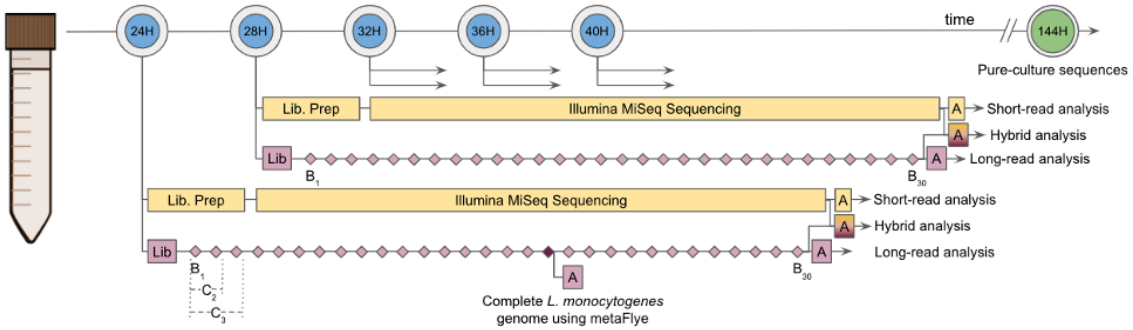


Figure 5.1: The effective time required to sequence and analyze the quasimetagenomic samples. The blue circles marked as 24H, 28H, 32H, 36H, and 40H denote the five enrichment time points where the quasimetagenomic samples were collected and sequenced with the Illumina MiSeq (short read) and the Oxford Nanopore GridIon (long read). Diamonds represent the 30 batches (B_1 to B_{30}) of 4000 GridIon reads, each generated 45 min apart. For our analysis, reads from each batch were merged with previously obtained batches to form cumulative batches (C_i). The time taken to assemble the reads is shown with boxes labeled 'A'. C_{18} at 24H marks the earliest time point where a complete *Listeria monocytogenes* genome was reconstructed (with `metaFlye`). The green circle corresponds to the time required to culture and sequence a pure colony isolate of *Listeria monocytogenes* i.e. 144H. *Note: bioinformatic analysis can be performed in "real-time" on the GridIon batches as they are output whereas an Illumina MiSeq sequencing run must finish before the bioinformatics can begin. However, for our analysis we partitioned the reads from each MiSeq run into 30 batches—each composed of an equal number of sequenced bases as the GridIon batches*

The mean read length for C_{30} across enrichment time points ranged from 174 to 198 nucleotides for Illumina MiSeq and 1,923 to 4,445 nucleotides for Oxford

	24H	28H	32H	36H	40H
Sequenced base pairs	2.3×10^8	3.3×10^8	3.0×10^8	5.4×10^8	5.0×10^8
#GridIon reads	1.2×10^5	1.2×10^5	1.2×10^5	1.2×10^5	1.2×10^5
MiSeq reads in C_{30} (total reads)	1.2×10^6 (2.9×10^6)	1.9×10^6 (4.0×10^6)	2.2×10^6 (3.6×10^6)	3.0×10^6 (3.5×10^6)	2.7×10^6 (2.9×10^6)

Table 5.1: Summary of sequence data for C_{30} at each enrichment time.

Nanopore GridIon. The longest sequenced GridIon read was 69,402 nucleotides long (Table 5.2). For the GridIon, there was a general increase in the mean and maximum read length as the enrichment time increased. Furthermore, the reads that mapped to the *L. monocytogenes* reference genome had a longer mean and maximum length compared to the rest of the reads across all enrichment time points (Supplementary Figure 1 from [99]). The putative *L. monocytogenes* reads also had a much lower mean GC content (38%) compared to the rest of the reads (49–54%) across enrichment time points (Supplementary Figure 2 from [99]).

Enrichment time	Mean read length	Max. read length	Average quality score	Min. est. sequencing error rate	Max. est. sequencing error rate
24H	1923	48588	21.8	7%	18%
28H	2721	55258	22.9	6%	17%
32H	3268	57233	22.8	7%	16%
36H	4445	62426	23.2	6%	13%
40H	4129	69402	23.2	6%	13%

Table 5.2: GridIon read length and sequencing error statistics for C_{30}.

The sequencing error rate for the reads mapping to the *L. monocytogenes* reference genome was 0.03% for the MiSeq reads and between 6.3 and 18% for the GridIon reads. The GridIon sequencing error rate has a range based upon whether the soft-clipping of read alignments (i.e. the ends of the reads not included in the

alignment range) was included as error or not. Each read is thus assigned two error estimates: an upper estimate of error that treats the unaligned portion of the read as an error, and a lower estimate that relies solely on the errors identified within the aligned range. Insertions, deletions, and mismatches were only counted for the aligned portion of the reads i.e. excluding the soft-clipped regions. For the long reads, 29.6%, 25.4%, and 45% of the errors were due to mismatches, insertions, and deletions, respectively—in accordance with previously published results [230]. For the MiSeq, the sequencing error rate and mean base quality were relatively uniform across samples. For the GridIon, the estimated sequencing error rate range decreased from 24H (7% to 18%) to 40H (6.3% to 13%) while the mean per-base quality score slightly increased over the same time period, from 21.83 to 23.19, respectively.

5.2.2 Selection of the reference genome

The accuracy of the assemblies was assessed with respect to a complete reference genome that had been isolated and sequenced (PacBio SMRT technology) from ice cream samples from the same facility as used for our analysis [258]. The reference was treated as a "gold standard" with an expected accuracy of ∼99.999% [259]. Previous research had shown that the outbreak consisted of two strains—one was isolated from Facility 1 only while the other was mainly isolated from Facility 2 [258]. The ice cream samples used for our analysis came from Facility 1. The reference genome used here had been used as a reference for SNP analysis of the

isolates from Facility 1, showing they differed by 29 SNPs or fewer. Another reference genome, from Facility 2, had been used as the representative of the second strain. The C_{30} `MegaHit` quasimetagenome assemblies showed a higher similarity with the reference from Facility 1 than Facility 2 (mean `Mash` [260] distance: 0.0206 and 0.0218 respectively). The reference from Facility 1 was subsequently used for our analysis.

The *L. monocytogenes* contigs derived from the quasimetagenomes were assessed for similarity to the reference sequence, and 55 loci were identified (46 single nucleotide insertions, 2 di-nucleotide insertions, and 5 SNPs) that differed at all enrichment times. Four of these variants (1 SNP and 3 single nucleotide insertions) occurred within the core of the *L. monocytogenes* genome (see Methods for a description of how the core was defined).

5.2.3 Assessing the presence of multiple *L. monocytogenes* strains

The presence of multiple, closely-related *L. monocytogenes* strains in the quasimetagenomes could affect the accuracy of the assemblies. A prior analysis of the ice cream samples [236] had identified three putative co-occurring *L. monocytogenes* strains based upon the detection of three 16S rRNA gene variants. However, analysis of the 16S rRNA genes in the reference genome identified 6 copies of the 16S rRNA operon which clustered, by sequence, within three distinct clusters consistent with the originally-determined variants.

The presence of multiple strains in the quasimetagenomes was assessed and 586 loci were identified (75 within the core genes) where the pile-up of MiSeq reads indicated the presence of two alleles, i.e. the reference allele and a variant. The percent of reads supporting the variants had a normal distribution with a mean of 17% and a standard deviation of 4%—indicating a 5:1 ratio of relative abundance. This evidence suggests that two highly-clonal strains co-occur in our quasimetagenomic samples.

5.2.4 General quasimetagenome assembly statistics

Ten assembly approaches were tested (Table 5.3), which were grouped into four broad categories: short read, long read, short read hybrid and long read hybrid. For simplicity, a tool was defined as a hybrid assembly approach if it used both short and long reads whether it be short read assemblies that get scaffolded with long reads (short read hybrid) or long read assemblies that get polished with short reads (long read hybrid).

All assembly approaches had a mean runtime (for full set of reads, C_{30}, across enrichment times) of approximately 40 min or less (Table 5.4) except `Canu` which had a mean runtime of 98 min per sample. The fastest assembly approach was `Redbean` with a mean runtime of just one minute (Supplementary Figure 3: [99]).

The contiguity of the assemblies was measured using several metrics: the total

Tool	Application	Abbreviation
`MegaHit`	short read metagenome assembler	Short Read
`Redbean`	long read genome assembler	Long Read
`Canu`	long read genome assembler	Long Read
`metaFlye`	long read metagenome assembler	Long Read
`Racon`	polishing long read assemblies with long reads	Long Read
`HybridSpades`	hybrid genome assembler; short read assembly followed by long read scaffolding	Short Read Hybrid
`Opera-MS`	hybrid metagenome assembler; short read assembly followed by long read scaffolding either *de novo* or using reference genome	Short Read Hybrid
`ntEdit`	polishing long read assemblies with short reads	Long Read Hybrid
`Pilon`	polishing long read assemblies with short reads	Long Read Hybrid

Table 5.3: Tested ten assembly approaches.

assembly length (Supplementary Figure 4 from [99]), number of contigs (Supplementary Figure 5 from [99]), N50 (Supplementary Figure 6 from [99]), and longest contig assembled (Supplementary Figure 7 from [99]). The mean values for C_{30} across enrichment times for each contiguity metric are described in Table 5.4. Approaches that first assemble short reads (short read and short read hybrid assemblies) contrasted substantially with those that first assemble long reads (long read and long read hybrid assemblies) having consistently longer total assembly lengths, orders of magnitude more contigs, lower N50s, and shorter longest contigs. In general, as the enrichment of *L. monocytogenes* progressed, there was a general decrease in the number of contigs and total assembly size (Supplementary Figures 4 and 5: [99]).

As expected, the long read and long read hybrid assemblies had the highest

Assembly tool	Runtime (min)	Total assembly length	No. of contigs	N50	Longest contig
`metaFlye` (Long Read)	40.6	4,291,417	27	3,056,133	3,056,133
`Canu` (Long Read)	98	3,470,967	21	1,754,979	2,071,553
`Redbean` (Long Read)	1	3,474,503	35	2,123,769	2,131,343
`MegaHit` (Short Read)	32.8	7,972,605	7,315	97,577	672,182
`metaFlye+Racon` (Long Read)	41.6	4,261,624	27	3,039,238	3,039,238
`HybridSpades` (Short Read Hybrid)	22.6	11,681,048	19,285	112,850	686,270
`Opera-MS` (no reference) (Short Read Hybrid)	12.2	10,340,211	13,921	105,382	655,220
`Opera-MS` (with reference) (Short Read Hybrid)	13.6	10,363,273	13,913	205,943	1,919,416
`metaFlye+Racon+Pilon` (Long Read Hybrid)	41.6	4,271,759	27	3,041,086	3,041,086
`metaFlye+Racon+ntEdit` (Long Read Hybrid)	41.6	4,274,358	27	3,041,440	3,041,440

Table 5.4: Mean assembly statistics (C_{30} at each enrichment time) for each assembly approach.

N50 values and the longest contigs—often near the reference genome length for *L. monocytogenes* (∼3 Mbp). Amongst the long read assembly tools, the metagenome assembler `metaFlye` consistently produced highest N50 values with longest contigs nearest to the length of *L. monocytogenes* reference genome (Table 5.4); however, the differences between long read assembly tools decreased with enrichment.

In contrast, the short read and short read hybrid assemblies had low N50 values and the longest contigs assembled were consistently shorter (often by orders of magnitude) with little to no increase beyond 60× depth of coverage. `Opera-MS`, using reference-guided scaffolding, was the main exception, assembling contigs of 2 Mbp or more at all enrichment time points.

5.2.5 Taxonomic composition of the quasimetagenomic samples

The number of species identified in the assemblies ranged from 2 to 10 with the short read and short read hybrid assemblies containing more species than the long read and long read hybrid assemblies (Figure 5.2). The number of species decreased with enrichment time, and *L. monocytogenes* and *Rothia mucilaginosa* were the only species detected at all time points. *Bacillus cereus* was the most closely related species to *L. monocytogenes* detected in the quasimetagenomes (both species are members of the order *Bacillales*).

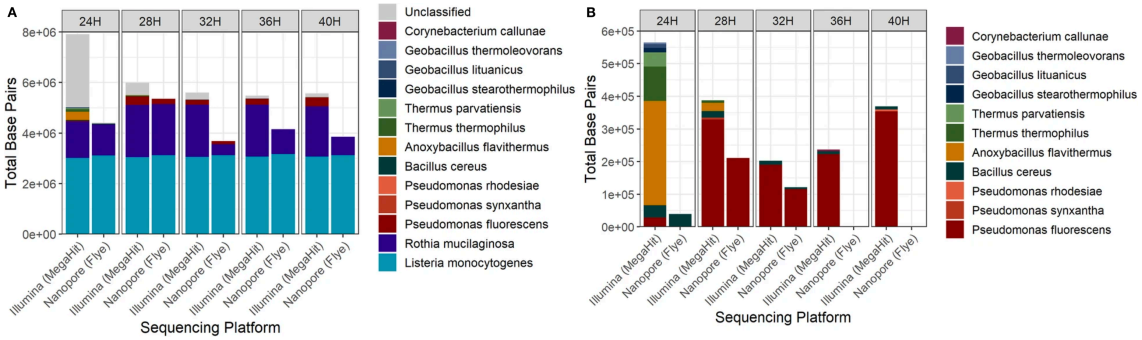


Figure 5.2: Taxonomic classification of C_{30} from each enrichment time point. For clarity, only the short read `MegaHit` and long read `metaFlye` assemblies were plotted (short read assembly results mirrored short read hybrid assemblies and long read assemblies mirrored long read hybrid assemblies). (A) The total bp of contigs per species (must have a minimum of 5000 bp) classified by Kraken. (B) Species in sample, excluding *L. monocytogenes*, *R. mucilaginosa* and unclassified sequences highlights how the short read assemblies capture more species than the long read assemblies.

L. monocytogenes was the most abundant species at all times and its abundance increased with enrichment time, but the abundance estimates differed for MiSeq and GridIon (Table 5.5). At 24H, 33% and 60% of MiSeq and GridIon reads, respectively, mapped to the *L. monocytogenes* reference genome. At 40H, 92% and

97% of MiSeq and GridIon reads, respectively, mapped to the reference genome.

Enrichment time	MiSeq (reads mapped with Bowtie2)	GridIon (reads mapped with MiniMap2)
24H	33%	60%
28H	68%	88%
32H	75%	94%
36H	88%	97%
40H	92%	97%

Table 5.5: Percent of reads that map to *L. monocytogenes* reference genome.

5.2.6 Reconstruction of *L. monocytogenes* from quasimetagenomes

The most contiguous recovery of the *L. monocytogenes* genome, as measured by the mean NG50 across enrichment time points (only using C_{30} at each time point), was by long read and long read hybrid assembly approaches (Figure 5.3). For the long read assemblers `Canu`, `Redbean`, and `metaFlye`, the mean NG50 values were 1,535,966 bp, 1,568,760 bp, and 2,490,733 bp, respectively. `metaFlye` assembled genome-length contigs for *L. monocytogenes* the most consistently among the long read assemblers—only `metaFlye` assemblies were used for long read hybrid assemblies. The long read hybrid approaches (using `metaFlye` and `Racon` in combination with `Pilon` or `ntEdit`) slightly decreased the mean NG50 of the `metaFlye` assemblies, 2,477,272 bp, 2,478,715 bp, 2,478,772 bp, respectively.

The short read `MegaHit` assemblies had the smallest mean NG50 at 162,346 bp. The short read hybrid assemblies of `HybridSpades` and `Opera-MS` without reference-guided scaffolding had mean NG50's that were several fold higher than

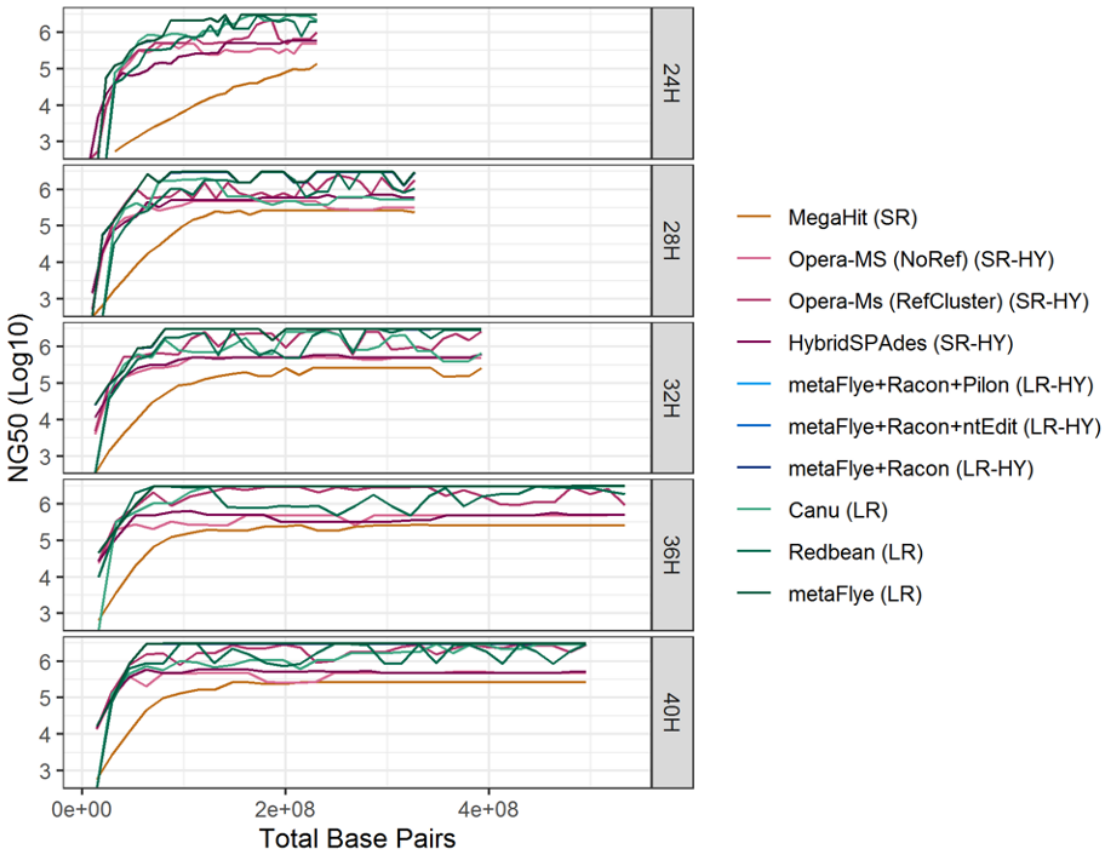


Figure 5.3: The NG50 versus the total number of base pairs sequenced per cumulative batch for the assembled *L. monocytogenes* contigs at each of the enrichment time points for each assembly approach. (Abbreviations: SR=short read, LR=long read, HY=hybrid)

the `MegaHit` assemblies, 431,211 bp and 375,881 bp, respectively. `Opera-MS`, using reference-guided scaffolding, had a mean NG50 of 1,414,301 bp, nearly an order of magnitude higher than `MegaHit` and close to that of the long read assembler `Canu`.

Only the long read assemblers were able to assemble genome-length contigs (over 3 million bp) for *L. monocytogenes*. The earliest complete reconstruction of the *L. monocytogenes* genome was at 24H and C_{18} with `metaFlye` (33× depth of coverage of the *L. monocytogenes* genome), 24H and C_{22} with `Canu` (40× depth of coverage of the *L. monocytogenes* genome), and 28H and C_{16} with `Redbean` (47× depth of

coverage of the *L. monocytogenes* genome). The genome length contigs, irrespective of the long read assembly approach, were frequently up to tens of thousands of base pairs longer than the reference genome, mainly due to over-circularization of the assembly by a read length or less. Additionally, each long read assembler recovered a circularized 71 kbp putative *L. monocytogenes* plasmid that was always fragmented in the short read assemblies. The best BLAST hits within the NCBI nt database for the assembled plasmid were to known *L. monocytogenes* plasmids (NCBI accessions CP053631.1 and CP044431.1). The plasmid was not found to host any known resistance or virulence genes.

5.2.7 Assembly errors in *L. monocytogenes* genomes reconstructed from quasimetagenomes

`Quast` was used to compare the mean number of misassemblies, mismatches per 100 Kbp, and indels (insertions and deletions) per 100 Kbp in the *L. monocytogenes* contigs for each assembly approach, given the highest sequencing depth of coverage of the quasimetagenomes (i.e. C_{30}) across enrichment times (Figure 5.4). The number of misassemblies and mismatches varied more by tool than assembly strategy. The mean number of misassemblies ranged from 10.8 (`Canu`) to 0 (`HybridSpades`). The mean number of mismatches per 100 Kbp ranged from 31.8 (`Redbean`) to 1.2 (`metaFlye`). In contrast, the long read assembly approaches had a pronounced indel rate versus other approaches, ranging from 265 (`Canu`) to 481

(`metaFlye`). The combination of `metaFlye` with `Racon` substantially reduced the number of indels to 74 per 100 kbp. Combining short read and long read information with long read hybrid assembly approaches further reduced the number of indels to $\sim$ 3 per 100 kbp. Short read assembly/short read hybrid assembly approaches had the lowest indel rate of around 1 to 2 per 100 kbp.

5.2.8 Accuracy of *L. monocytogenes* metagenome-assembled genomes

At all enrichment time points and C_{30} reads (for both short and long reads), there was 100% breadth of coverage of the *L. monocytogenes* reference genome and up to $\sim 160\times$ depth of coverage.

The fraction of the *L. monocytogenes* genome that was typeable by the MiSeq and GridIon reads was assessed by identifying regions in the reference genome where the C_{30} reads mapped ambiguously (i.e. mapped with the same alignment score to multiple genome locations). For the MiSeq and GridIon reads, a median of 3.9% (118,615 bp) and 0% (0 bp), respectively, of the reference genome consisted of ambiguous regions.

Earlier results provided evidence for the presence (with a 5:1 relative abundance ratio) of two strains of *L. monocytogenes* in the quasimetagenomes. The less abundant strain differed from the more abundant strain at 586 loci. Analysis with `Snippy` showed that no more than 13 of the 586 variants in the low abundance strain

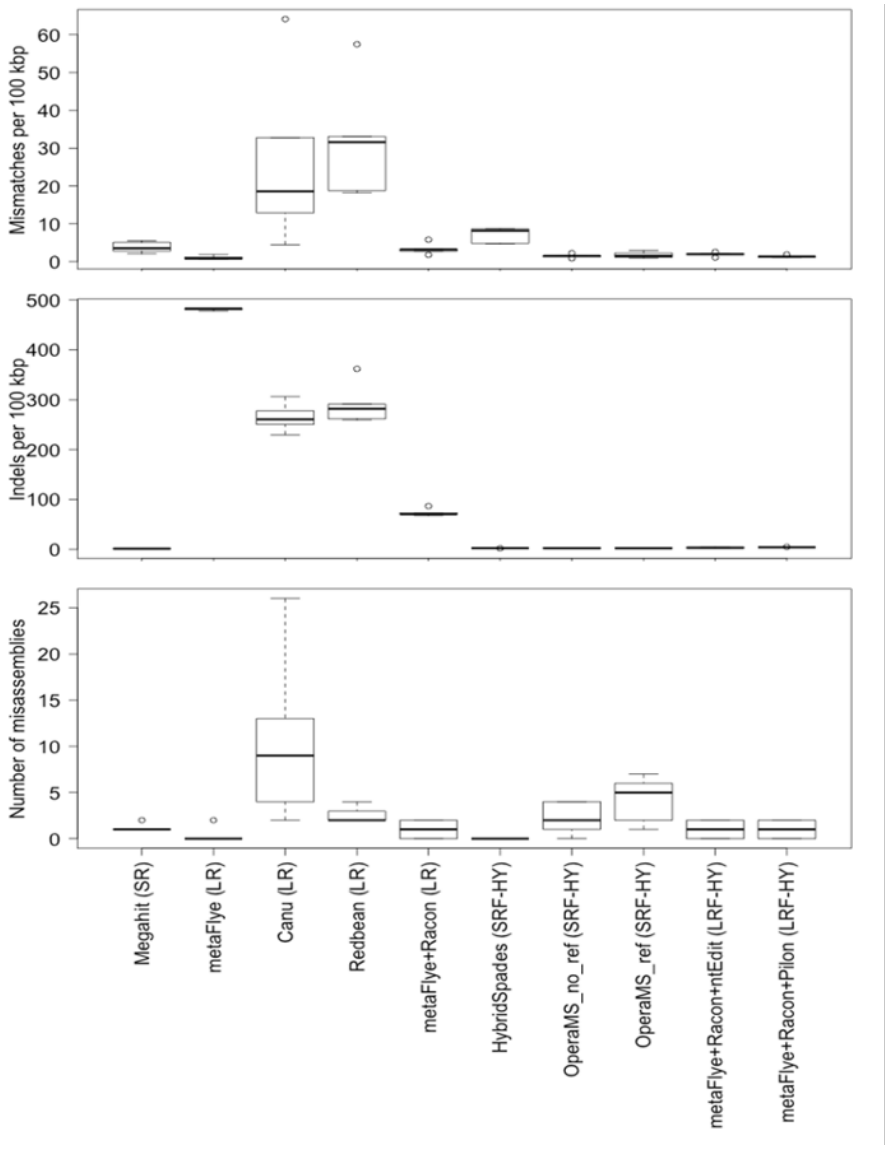


Figure 5.4: The quality of assembled contigs annotated as *L. monocytogenes*, with respect to the reference genome, using `Quast` for C_{30} at each of the enrichment time points. The number of mismatches, insertion/deletion (indels), and misassemblies per 100 kbp for each assembly approach. (Abbreviations: SR=short read, LR=long read, HY=hybrid)

were present in a given C_{30} assembly (across enrichment times). However, the long read assemblies contained the highest median number of variants (maximum was 12 with `metaFlye`) while the other assembly approaches had a median of 3 or less.

Next, the accuracy of the assemblies (C_{30}–C_{30} at each enrichment time) was

assessed by calculating the BLAST distance between the core genes (Figure 5.5) and the complete set of genes (Figure 5.6) of the reference genome and the *L. monocytogenes* contigs. As defined earlier, the BLAST distance is a measure of sequence similarity equalling the number of mismatches, insertions, and deletions in the BLAST alignment between the reference genes and the assembled genes. The short read and short read hybrid assemblies attained the smallest BLAST distances for the core genes, while the long read hybrid assemblies attained the smallest BLAST distances for the complete set of genes.

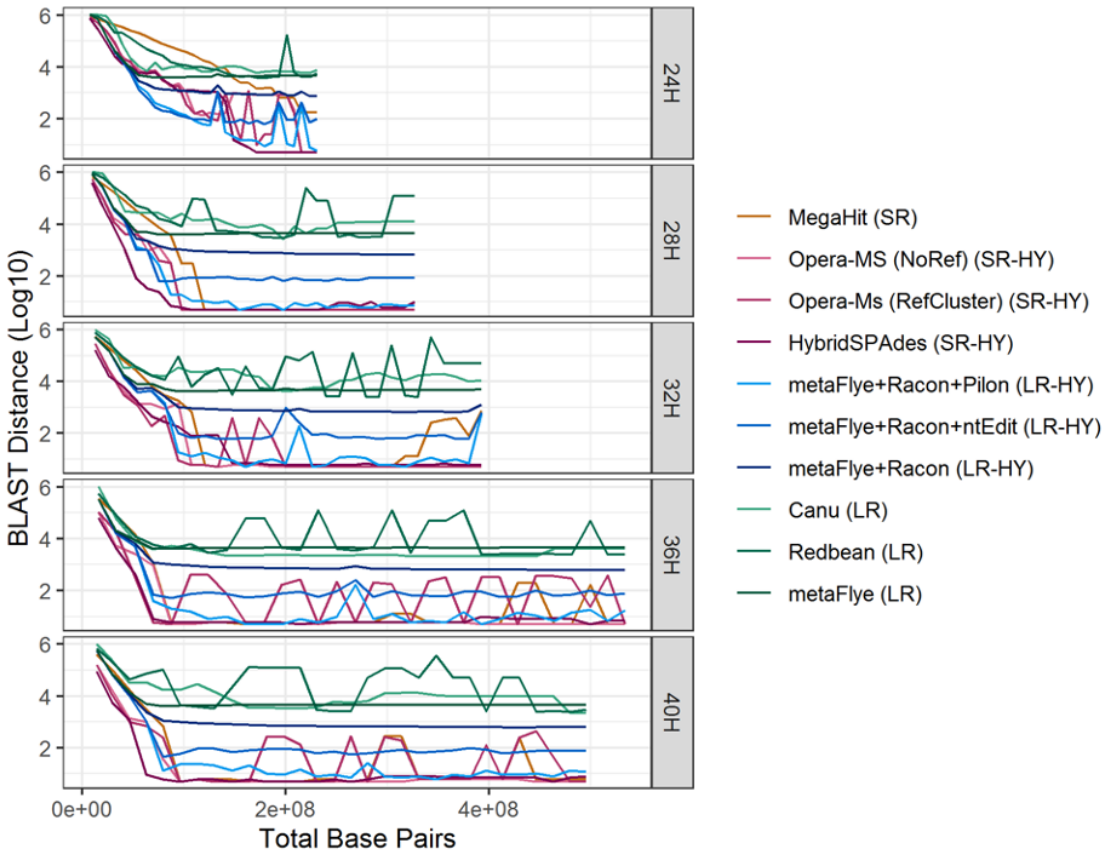


Figure 5.5: Core gene BLAST distances. BLAST distance between the core genes of the reference genome and the assemblies versus the total number of base pairs sequenced per cumulative batch. (Abbreviations: SR=short read, LR=long read, HY=hybrid)

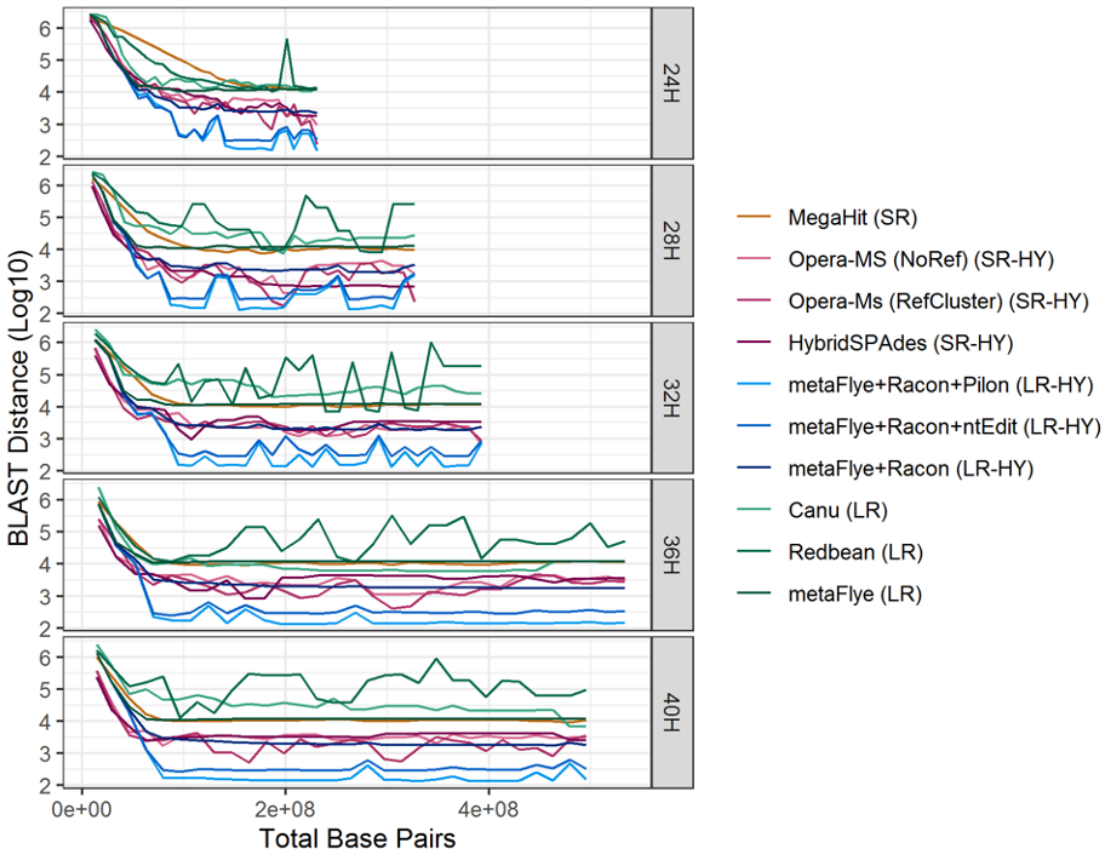


Figure 5.6: Complete gene set `BLAST` distances. `BLAST` distance between the complete gene set of the reference genome and the assemblies versus the total number of base pairs sequenced per cumulative batch. (Abbreviations: SR=short read, LR=long read, HY=hybrid)

For the core genes, the smallest `BLAST` distance observed was 5 (Figure 5.5). Four of the differences were caused by variants identified previously in the core genes of the *L. monocytogenes* extracted from the quasimetagenomes. The fifth difference varied in location for different assemblies, and showed no relation to the variants discovered previously.

Short read hybrid approaches assembled the core genes with `BLAST` distance 5 at the earliest time point: `HybridSpades` at 24H and C_{22} corresponding to 40× (long reads) and 19× (short reads) depth of coverage of *L. monocytogenes* reference

genome; Opera-MS, both with and without reference-guided scaffolding, at 24H and C_{28} corresponding to 50× (long reads) and 25× (short reads) depth of coverage of *L. monocytogenes* reference genome. MegaHit assemblies attained a BLAST distance of 5 after 28H and C_{11} corresponding to 28× depth of coverage of *L. monocytogenes* reference genome. At 24H, 28H and 36H the short read hybrid assemblies obtained a BLAST distance 5 with fewer short reads than the short read assemblies; however, at 32H and 40H, the short read and short read hybrid assemblies required the same amount of short read data to achieve a BLAST distance of 5.

The long read assemblies never achieved a BLAST distance of less than 2000 even with 158X depth of coverage of *L. monocytogenes*. Polishing the long read metaFlye assemblies with Racon improved the assembly of the core genes, achieving a minimum BLAST distance of 609. Long read hybrid assembly with Pilon achieved a BLAST distance of 5 at 28H and C_{14} which corresponded to 36X (short reads) and 38X (long reads) depth of coverage of *L. monocytogenes* reference genome; however, it achieved BLAST distance 5 less consistently than short read or short read hybrid approaches (Figure 5.5). Long read hybrid assembly with ntEdit assembled the core genes with less accuracy than Pilon, with a median BLAST distance (C_1–C_{30} across enrichment times) of 81 and 11, respectively.

The long read hybrid approaches assembled the complete gene set with the lowest BLAST distance, with Pilon outperforming ntEdit (Figure 5.6). Pilon achieved a BLAST distance of 132, the best observed for any tool, at 28H and C_{14} corresponding to 36× (short reads) and 38× (long reads) depth of coverage of the *L.*

monocytogenes reference genome. The mean `BLAST` distance across enrichment time points was 699 for `Pilon` and 798 for `ntEdit`. None of the other assembly approaches attained this level of accuracy. For reference, the next best tool, `metaFlye+Racon`, had a mean `BLAST` distance of 2991.

5.2.9 Variation in assembly quality between successive cumulative batches

In addition to accuracy, the precision with which assemblies can be reconstructed is of great importance for pathogen detection. The accuracy of assembly approaches (in terms of divergence in core and full gene sets with respect to the *L. monocytogenes* reference) varied widely between successive cumulative batches (Figure 5.7).

`Opera-MS` (without reference guided scaffolding), `HybridSpades`, `MegaHit`, and `metaFlye+Racon+Pilon` most consistently assembled the core genes; the median difference in `BLAST` distance between successive cumulative batches for these tools was 0, 1, 1, and 5, respectively, across all enrichment time points. All assembly approaches had a median difference in `BLAST` distance of less than 50 except `Opera-MS` with reference guided scaffolding (121), `Canu` (1183) and `Redbean` (10,930).

The variability in the accuracy for the reconstruction of the complete gene set was an order of magnitude greater than for the core genes. The most consis-

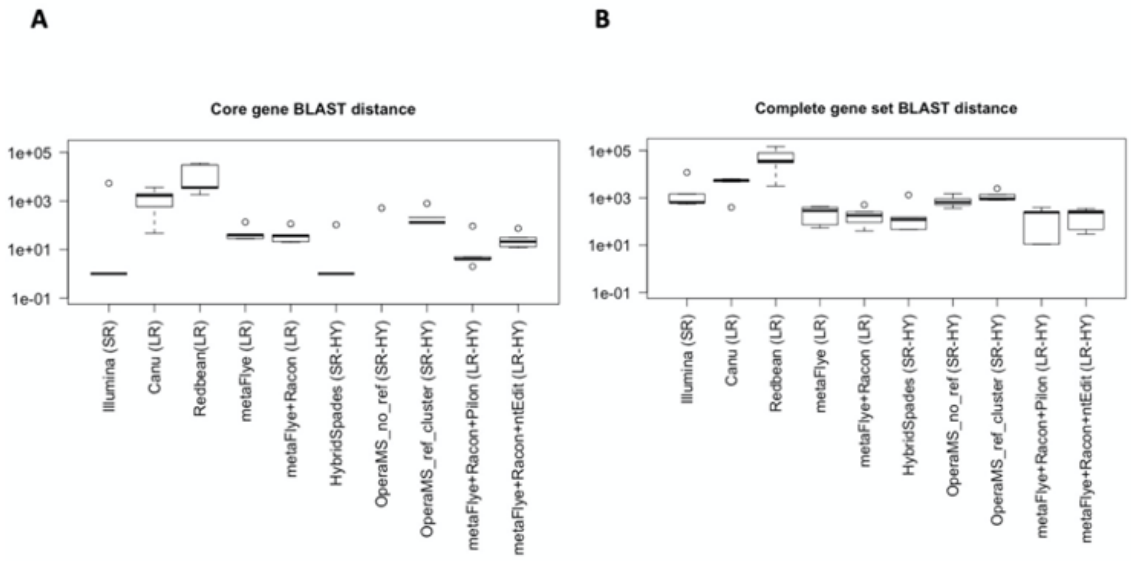


Figure 5.7: Consistency of assembly approaches between successive cumulative batches. Median successive cumulative batch difference in `BLAST` distances, across enrichment time points, for the (A) core genes and (B) complete genes (Abbreviations: SR=short read, LR=long read, HY=hybrid)

tent tools were `HybridSpades`, `metaFlye+Racon+Pilon`, `metaFlye+Racon+ntEdit`, `metaFlye+Racon`, and `metaFlye`—the median difference in BLAST distance between cumulative batches was 132, 137, 140, 183, and 207, respectively. All assembly approaches had a median difference in `BLAST` distance of less than 1000, with the exceptions of `Opera-MS` with reference guided scaffolding (1019), `Canu` (4758), and `Redbean` (32,655).

5.2.10 Depth of coverage did not always improve assembly quality

Increased depth of coverage did not always correlate with improved performance in assembly metrics. For example, the longest contig assembled by the short

read assemblies was very similar at 30× (695,760 nt) and at 150× (695,778 nt) depth of coverage. In some cases, the performance of assembly approaches actually decreased with increased depth of coverage. For example, the lowest `BLAST` distance for the complete gene set for the `metaFlye+Racon+Pilon` assemblies increased from 132 to 153 despite an increase of 100× depth of coverage of the *L. monocytogenes* genome for both short and long reads.

5.3 Discussion

Public health labs are continually developing and testing new methods and approaches to increase the speed and resolution of pathogen source tracking. Expediting source attribution will contribute to reduced illnesses, deaths and the economic burden of illness outbreaks. Currently, the standard workflow for strain typing and source attribution involves sequencing genomes (primarily with Illumina MiSeq technology) of isolated colonies, cultured from selective enrichments. Sequence data is analyzed using SNP and/or MLST analyses. Here we evaluated the contribution of quasimetagenomics and the applied integration of (short) MiSeq and (long) GridIon reads for the improvement of this workflow.

5.3.1 Quasimetagenomics expedites source tracking

Currently, direct metagenomic sequencing of samples cannot replace genome sequencing of culture isolates for the strain typing of pathogens; however, quasimetagenomics has shown great promise for reducing the amount of enrichment time needed to type pathogens with sequence data [75, 80, 248]. Previous work on the listeriosis ice cream outbreak demonstrated that quasimetagenomic short read sequencing provided sufficient coverage of the *L. monocytogenes* genome to determine its membership in the outbreak cluster at 24H enrichment—a significant improvement over the $\sim$ 6 day procedure required to culture and sequence an isolate genome [80]. This work supports that MiSeq short read sequencing can expedite the recovery of a target pathogen from quasimetagenomes, accurately reconstructing the *L. monocytogenes* core genes at 28H of enrichment. Further, the integration of MiSeq short read and GridIon long read sequencing further expedited the accurate assembly of the core genes and increased the contiguity of assemblies—including the reconstruction of a complete genome and plasmid—at 24H of enrichment (Figure 5.1). This highlights that an integrated approach to quasimetagenomics can greatly expedite and enhance source tracking.

5.3.2 Long reads have added value over short reads for quasimetagenomics

Although short reads can be used for high resolution SNP and cgMLST/wgMLST analyses they cannot span many genomic repeat regions, resulting in fragmented assemblies that preclude the recovery of complete bacterial genomes [79]. The fragmented assemblies can prevent the identification of genes, gene synteny, repeats, structural variants, and extrachromosomal sequences, like plasmids and phages, that could be readily observed in complete assemblies.

Our results showed that $\sim$ 4% of the *L. monocytogenes* genome was not typeable by the MiSeq reads. In contrast, the entire *L. monocytogenes* genome was typeable with the GridIon reads, enabling the complete reconstruction of the *L. monocytogenes* genome and plasmid at 24H of enrichment and only 33$\times$ depth of coverage. The ability of long reads to span genomic repeats will support much higher resolution whole genome based source tracking methods and provide detailed information about the mobileome. However, similar to previous studies [251], we found the high sequencing error rate of the nanopore reads to induce incorrect base calls in the assembled sequence, thereby negatively impacting strain typing and strain attribution. Nonetheless, with time, we expect the sequencing error rate to decrease and the utility of nanopore sequencing for source tracking to increase substantially.

Another advantage of nanopore over MiSeq sequencing is that the data is

output in batches of reads every 30 to 60 min (as opposed to a MiSeq sequencing run which takes $\sim$ 24H depending on the number of cycles). As assembling the reads is much faster than sequencing itself (Figure 5.1), nanopore sequencing allows the analysis to terminate as soon as sufficient reads have been obtained for accurate analysis—a point that may vary depending on the characteristics of the sample. This ability can greatly expedite source tracking by facilitating near-real-time bioinformatic analyses.

5.3.3 Hybrid assembly outperforms other approaches but with trade-offs

Our results support the accuracy of hybrid assemblies [75, 233, 234]—hybrid assembly, using both Illumina short reads and nanopore long reads, could reconstruct more complete and accurate genomes than using either of the platforms alone. However, the initial assembly strategy (i.e. whether the short reads were assembled first or long reads) had a substantial impact on the quality of the reconstructed genomes. Short read hybrid assembly approaches led to a more accurate assembly of the core genes, but the assemblies were more fragmented. The use of reference genomes to scaffold assemblies increased the contiguity of the short read hybrid assemblies, but also introduced assembly errors—a potential consequence if the references used for scaffolding has structural differences compared with the genomes being assembled. For the long read hybrid assembly approaches, a higher indel

rate prevented the accurate assembly of the core genes; however, the assemblies had higher contiguity, sometimes reconstructing the complete *L. monocytogenes* genome. Additionally, the long read hybrid assembly approaches led to the most accurate recovery of the complete set of genes, with potential implications for characterizing the phenotype (e.g., drug resistance) of the pathogen. The choice of the hybrid assembly approach can be made subject to whether the application of the reconstructed genome mandates highly accurate core genes or an overall accurate complete genome.

5.3.4 Short read based assembly approaches showed the best performance

Assemblies need to be accurately reconstructed to be useful for SNP and cgMLST/wgMLST based source tracking analyses. Among the assembly approaches tested, the most accurate was the reconstruction of the core genes using either the short read or short read hybrid assembly strategy. Short read hybrid assembly was consistently able to accurately assemble the core genes with the same amount or fewer short reads than the short read assemblies. However, the combined use of short and long reads entails higher costs in both personnel time and reagents, which may not be justified as similar accuracy can be obtained with short reads alone at a slightly higher depth of coverage. In contrast, no assembly approach could reconstruct the complete set of genes with high accuracy or consistency, although

long read hybrid approaches were by far the best performing. Nonetheless, given a lower sequencing error rate, long read approaches might become preferable with the added value of assembling complete genomes and mobile elements like plasmids.

5.3.5 Areas of improvement for assembly algorithms

At the time of our analysis, `metaFlye` was the only metagenomic long read assembler available, and it performed better than the other long read assemblers in our application. This observation highlights that long read assemblers developed for single genomes are not effective when samples contain mixtures of DNA from multiple organisms—suggesting the need for further research in developing efficient metagenomic assembly tools for long read data. Additionally, the quality of the `metaFlye` assemblies was improved considerably by polishing the assemblies with the long reads themselves, indicating that none of the long read assemblers make full use of the information available in the long reads.

The observed differences between hybrid assembly approaches that start with short reads and those that start with long reads, suggest that hybrid approaches are currently limited by the weaknesses of the different technologies. This highlights the scope for improvement in hybrid assembly approaches, underscoring that we are still far from developing techniques which effectively integrate their strengths (e.g., the contiguity of long reads and high per-base quality of the short reads).

A weakness common to all assembly approaches was sensitivity to the addition of cumulative batches of sequence data, resulting in inconsistent gains/losses in assembly quality. This affected many metrics such as the N50 and the accuracy of the assembled genes. The differential sensitivity of assembly approaches to the addition of sequence data from the same sample suggests that assembly tools can be made more robust and consistent—greatly benefiting many applications including strain typing.

An advantage for quasimetagenomics is the detection of co-occurring strains that might be missed by traditional methods (i.e. culturing and sequencing a single isolate). Our analysis suggested the presence of at least two strains of *L. monocytogenes* in the quasimetagenomes. However, the current tools do not account for the variations within the (quasi-)metagenomic samples and current assembly approaches simply reconstruct the most abundant strain, which is what we observed with our assemblies. While further analysis of the data can reveal the strain structure hidden by the consensus assembly, we believe it is preferable that assemblers themselves account for and reveal the strains contained in the sample, information that could be valuable for source tracking.

5.4 Conclusion

The integration of nanopore long read and Illumina short read sequencing expedited the reconstruction of high quality *L. monocytogenes* assemblies from

ice cream quasimetagenomes. The core genes were accurately reconstructed after 24H enrichment with the short read hybrid assemblies and 28H for the short read assemblies—a significant reduction from the standard 6 day protocol. Although the GridIon long read assemblies had too many errors to reconstruct the core genes with high fidelity, they had added value for reconstructing complete genomes and plasmids—providing information about synteny, gene content and genome structure that were not accessible with short reads. Hybrid assembly showed the best performance but with different weaknesses depending on whether the short or long reads built the initial assembly—highlighting areas for algorithmic improvement that integrate the strengths of long and short reads (e.g. the contiguity of long reads and high per-base quality of the short reads). A new and more complete level of information about genome structure, gene order and mobile elements can be added to the public health response by integrating microbiological (quasimetagenomic), molecular (long and short read sequencing) and optimized bioinformatic approaches.

5.5 Methods

5.5.1 Experimental design

Using long and short read sequencing technologies, we compared the performance of various assembly approaches for reconstructing the genome of *L. monocytogenes* from selective enrichments of naturally contaminated ice cream samples

(Figure 5.1). The isolation of a pure colony of *L. monocytogenes* for sequencing typically requires up to 6 days of selective culture enrichment [80]. During the selective enrichment, aliquots were collected at 4-hour intervals from 24 to 40 hours (denoted as 24H, 28H, 32H, 36H, 40H). MiSeq short read and GridIon long read sequencing were performed on DNA from these incremental enrichments. At each time point, over a range of sequenced depth of coverage of the quasimetagenomes, the sequence data was assembled using the short and long reads in combination and separately. Assembly quality was evaluated by comparison to a complete *L. monocytogenes* reference genome—sequenced and assembled from PacBio data—obtained from the full 6-day enrichment protocol.

5.5.2 Enrichment

Ice cream samples, associated with the 2015 Blue Bell multistate listeriosis outbreak, were homogenized and added to Buffered Listeria Enrichment Broth (BLEB) with pyruvate according to the specifications outlined in Chapter 10 of the FDA BAM [235]. The mean MPN/g of *L. monocytogenes* in the ice cream samples was 11.99. After four hours, three filter sterilized selective agents (M52) were added to achieve final concentrations of 10 mg/L acriflavin, 40 mg/L cycloheximide, and 50 mg/L sodium nalidixic acid in the BLEB. Four replicates of negative (no ice-cream) and positive controls (*L. monocytogenes* cells) were also evaluated for bacterial growth every four hours over the 40H enrichment.

5.5.3 DNA extraction and sequencing for short reads

For each of the enrichment time points (24H, 28H, 32H, 36H, and 40H), DNA was extracted using DNeasy Blood and Tissue kit (Qiagen) following the protocol for Gram-positive bacteria with minor modifications: 1.5 ml of the culture was pelleted (5000×g, 15 min) and the pellet resuspended in 200 mL of enzymatic lysis buffer containing 20 mM Tris-HCl (pH -8.0), 2 mM Sodium EDTA, 1.2% Triton X- 100, 20 mg/ml of lysozyme. Samples were incubated for 60 min at 37 °C. Short read libraries were prepared with Nextera Flex (Illumina) library prep kit according to the manufacturer's specifications. Libraries from enrichment time points 24H, 28H, 32H, 36H, and 40H were multiplexed along with 20 other libraries from different time points from the same study on to Illumina MiSeq 2×250 cartridge (Illumina, CA) following manufacturer recommended protocol.

5.5.4 DNA extraction and sequencing for long reads

For each enrichment time point (24H, 28H, 32H, 36H, 40H), 2 ml aliquots of enrichment were removed and pelletized using a benchtop Centrifuge (Eppendorf 5418 R, NY, USA) at 4000 rpm for 10 mins. The pellet was resuspended in 300 μl of TE Buffer. 300 μl of the resuspended cells were loaded on the Maxwell® RSC Instrument (automated DNA extraction instrument, Madison, WI, USA) cartridge for DNA extraction. Genomic DNA was extracted using Maxwell® RSC Cultured

Cells DNA Kit (Cat no: AS1260, Madison, WI, USA) on Maxwell RSC instrument following the manufacturer recommended protocol for Gram-positive bacteria.

Sequencing libraries were prepared using the ligation sequencing kit (Cat no: SQK-LSK109, Oxford Nanopore, Oxford, UK), according to the manufacturer's specifications along with Native Barcoding Expansion 1–12 (Cat no: EXP-NBD104, Oxford Nanopore, Oxford, UK) for multiplexing the samples. The libraries were multiplexed into 2 pools (Pool1: 24H, 28H, 32H): Pool 2: 36H, 40H). The libraries were sequenced using GridIon with Flow cell (Cat no: FLO-MIN106, Oxford, UK) following the manufacturer's recommended protocol. The GridIon outputs the raw signal data in batches of 4000 sequenced reads in fast5 format files. Each fast5 file was converted into fastq formatted DNA sequences using `Guppy` for basecalling. The fast base calling mode was used, which has a speed of $\sim$ 4.6 Mbp/second. The GridIon typically outputs a batch of reads every 30 to 60 min (internal to lab), but is affected by factors such as the length and quality of the DNA fragments being sequenced.

5.5.5 *L. monocytogenes* reference genome

Previous work identified two strains from the *L. monocytogenes* ice cream outbreak [258]. Two reference genomes (NZ_CP016213.1 and NZ_MAGN00000000.1) were used for the SNP analysis of the two strains. These reference genomes were compared with the *L. monocytogenes* assemblies from the quasimetagenomes using

`Mash` (v2.0, $k = 25$, $s = 100,000$). The complete *L. monocytogenes* genome (Genbank accession NZ_CP016213.1) was more similar to the data in the quasimetagenomes (see Results section) and was used as the reference for our analyses. This reference organism had previously been isolated from a single colony at the end of the enrichment protocol and sequenced with PacBio RSII from ice cream samples from Facility 1, the same facility our samples came from [258]. The reference is 3,030,827 bp long with 2984 protein-coding genes (2,710,041 bp in total length) predicted by `Prokka` (v1.12) [147] and a GC-content of 38%. The core genes (using the 1013 gene cgMLST scheme developed for *L. monocytogenes* at the FDA [223, 224]) were identified by `BLAST` [156] alignment. The total length of the core genes was 1,075,554 bp. Six copies of the 16S rRNA were identified in the reference genome with `BLAST` using the `RNAmmer` database [261].

5.5.6 Partitioning the sequenced reads into cumulative batches

The GridIon nanopore sequencing instrument generates the data in batches of 4000 reads, denoted here as B_n for the n^{th} batch. Our analysis used the first 30 batches of reads, i.e. the first 120,000 reads corresponding to batches $B_1, B_2, \ldots, B_{30}$ (Figure 5.1). To analyze the quality of assemblies as a function of increased sequencing depth, each successive batch of reads was combined with the previous batches for assembly to form "cumulative batches", denoted as $C_1, C_2, \ldots, C_{30}$, where $C_n = B_1 + B2 + \ldots + B_n$ (Figure 5.1). To compare assembly results strictly

based on sequencing technology, the number of base pairs for the MiSeq and GridIon data was normalized. Over a range of sequencing depths, MiSeq raw read files were partitioned into 30 corresponding batches of read pairs to match the cumulative batches by number of base pairs for GridIon reads. Table 5.1 records the total number of sequenced bases per C_{30} at each enrichment time.

5.5.7 Detection of genomic variants and the presence of multiple strains

The detection of variants between the reference and the *L. monocytogenes* sequences reconstructed from the quasimetagenomes was conducted with two methods. In both cases, the MiSeq reads from cumulative batch C_{30} from each enrichment time were analyzed. The first method called variants with `Snippy` (v4.6) if there was 10X depth of coverage and $\geq 95\%$ of the reads supported the variant. The second method consisted of mapping the MiSeq reads to the reference genome with `Bowtie2` (v2.3.4) [262] and analyzing the pile-up of reads with `SAMtools` (v1.7) [263]. Loci with $\geq 50\times$ depth of coverage and where 20 to 90% of the aligned reads indicated the presence of another allele (while the rest of the aligned reads supported the reference allele) were considered to be evidence for multiple strains.

5.5.8 Raw read statistics and reference genome coverage

Raw read statistics were collected for the 30 batches of reads (B_1–B_{30}) per enrichment time point, including: mean per base quality score, number of reads, number of base pairs, read length distribution, and estimated sequencing error rate. To estimate the sequencing error rate, the short and long reads were mapped to the *L. monocytogenes* reference genome with `Bowtie2` (v2.3.0) and `MiniMap2` (v2.17-r974-dirty) [264], respectively, using default settings. The number of mismatches, insertions, and deletions were counted for the mapped reads with respect to the reference genome. For the GridIon reads, an estimated range was provided for the sequencing error rate because `MiniMap2` is a local, as opposed to a global, read alignment tool. The range is based on whether soft-clipping of the read alignments is included as sequencing error (maximum estimate of error) or not (minimum estimate of error). Insertions, deletions, and mismatches were only counted for the aligned portion of the reads i.e. excluding the soft-clipped regions. The read mappings were used to estimate the breadth and depth of coverage (DOC) of the *L. monocytogenes* reference genome.

5.5.9 Assembling the sequenced reads

Short reads and long-reads from each cumulative batch (C_1–C_{30}) were assembled per enrichment time point (Figure 5.1). The short reads were assembled

using `MegaHit` (v1.2.9) [249] with default settings and scaffolded with `MetaCarvel` [152]. The long reads were assembled using `Canu` (v1.7) [252], `Redbean` (v2.5) [253], and `metaFlye` (v2.6-release) [250] with default settings. The `Redbean` assemblies were polished with `MiniMap2` (v2.17-r974-dirty) and `SAMtools` (v1.5) following the tutorial for `Redbean` on its GitHub page. Unlike `metaFlye`, which is a long read metagenome assembler, `Canu` and `Redbean` are not designed for metagenomic assembly. However, these assemblers were chosen for comparative analysis as they are frequently used long-read genome assemblers. All of the `metaFlye` assemblies were polished, using the long reads, with `Racon` (v1.4.15) [257]. `HybridSpades` (v3.14.0) [254] and `Opera-MS` (v0.8.3) [251] (with and without reference genome scaffolding) were used for short read hybrid assembly—short read assembly followed by scaffolding with the long reads. `Opera-MS` was chosen because it is a metagenome assembler, while `HybridSpades` was chosen because it is a popular genome assembler. `Pilon` (v1.23) [255] and `ntEdit` (v1.3.1) [256] were used for long read hybrid assemblies—long read assembly with `metaFlye` followed by short read polishing. Each tool was run with 12 cores of 2.70 GHz Intel Xeon E5–2680 processor.

5.5.10 Assembly statistics

The runtime (user time) of each assembly method on the server was recorded for cumulative batch C_{30} at each enrichment time point. `Quast` (v5.0.2) [145] was used to report the number of insertion/deletions/mismatches and the NG50 for the

C_{30} assembled *L. monocytogenes* contigs with respect to the reference genome.

General quasimetagenomic assembly statistics (total assembly length, the number of contigs, the longest contig, the N50) were collected for every cumulative batch (C_1–C_{30} at each enrichment time) using a custom Python script.

5.5.11 Comparison of the reference genome with *L. monocytogenes* assembled from the cumulative batches

We estimated the fraction of the reference genome where reads (MiSeq and GridIon) mapped ambiguously, i.e. mapped with the same alignment score to multiple genome locations. The MiSeq reads were mapped with `Bowtie2` and the GridIon reads were mapped with `MiniMap2`. The mean MAPQ score was calculated for each base of the reference genome. Loci with median scores lower than 40 were considered ambiguous [265].

The presence of alleles from the low abundance strain was assessed for the C_{30} assemblies (across enrichment times) with `Snippy` by aligning the assemblies to the reference genome and cross-referencing the variant loci identified when looking for multiple strains.

The *L. monocytogenes* contigs assembled from each cumulative batch (C_1–C_{30} at each enrichment time) were assessed for accuracy with respect to the reference genome. Accuracy was assessed by measuring the `BLAST` distance (a measure of

sequence similarity) between the predicted genes (both the core and complete set of genes) of the reference and the *L. monocytogenes* metagenome-assembled genomes. We define the `BLAST` distance as the number of mismatches, insertions, and deletions in the `BLAST` alignment between the reference genes and the assembled genes. Preferably, the edit distance between the reference genes and the genes found in the assemblies would have been calculated, but correctly identifying the entire length of genes, especially in noisy long read assemblies, is difficult; instead, the `BLAST` distance forms an approximation of the edit distance.

If the *L. monocytogenes* genome assembled from a sample comprised a single contig, the synteny of the core genes was compared to that in the reference.

5.5.12 Taxonomic classification

The contigs from the `MegaHit` (short read) assemblies and `metaFlye` (long read) assemblies (from each cumulative batch at every enrichment time) were taxonomically classified with `Kraken` (v1.1.1) [205] and the MiniKraken database using default settings. A species was considered present if $\geq$ 5000nt of contigs were annotated as that species.

The proportion of reads mapped to the *L. monocytogenes* genome was used as the relative abundance of *L. monocytogenes* in the samples.

Chapter 6: Detection of similar genomic regions shared between several metagenomic samples using SIMILE

As of writing this thesis, this chapter contains material under preparation for submission in ***SIMILE****: Discover similar genomic regions shared across a collection of metagenomic samples [266], which was a joint work with Hirak Sarkar, Hugh Rand, Rob Patro and Mihai Pop. KJ and MP designed the study. KJ developed the entire code, performed all bioinformatic analyses, created the figures, and wrote the manuscript. All authors analyzed the results and edited the manuscript.*

6.1 Introduction

With metagenomics, a microbial community can be sequenced directly from its environment; it has proven to be a key tool for the analysis of complex microbial communities, particularly the microbes that cannot be cultured in the laboratory using conventional isolation [81, 82]. Microbes and their genes from these complex communities are, together, referred to as the 'microbiome'; the microbiome plays an important role in human health, complex diseases, and environmental regula-

tion [267, 268, 269]. Whole shotgun metagenomics has facilitated the characterization of microbial diversity in the environment, identification of novel enzymes, bioprocesses, and natural products, steering the discovery of microbial applications in various industrial and biomedical settings [81, 83, 84]. Some of the applications where metagenomics has been increasingly adopted include food safety monitoring, spread of antimicrobial resistance factors, public health surveillance, and other comparative microbiome studies [85, 86, 87].

Whole shotgun metagenomics can facilitate microbiome analysis to explore the strain-level variations, structural variations and genomic regions corresponding to unclassified or understudied microbes, even compare the microbial compositions between samples from healthy individuals (controls) and diseased subjects (cases) in case-control microbiome studies [270, 271]. These microbiome comparisons rely on metagenomic assembly of microbial samples followed by the identification and analysis using genomic entities like SNPs, k-mers, genomes, taxonomic labels, and genes [88, 89]. Metagenome assemblies are, however, highly fragmented, comprising thousands of contigs; additionally, there is no 'ground truth' or *a priori* knowledge about the number of genomes present in the microbial sample or the genome associated with any contig [270]. Some microbiome studies have, therefore, relied on taxonomic identification and associating taxa presence-absence and relative abundances for case-control studies [85, 90]. However, the mere presence of a taxon is not always a deterministic factor for disease causality [91, 92, 93]. Some other approaches aim at contig 'binning' which attempt to group contigs by species' genomes or generate

clusters of contigs using genomic regions detected in multiple samples [272, 273]. Comparisons of multiple microbial samples using whole shotgun metagenomics facilitates the exploration of microbial diversity and identify genomic regions for biological post-processing and validation, such as obtaining taxonomic and functional profiles for corresponding genomic regions [270].

The comparison between multiple whole shotgun metagenomic samples, however, adds a layer of computational complexity to the already challenging problem of analyzing the unknown and uncharacterized microbial compositions [94]. One approach for metagenomic comparisons relies on co-assembly of multiple samples: it is, however, extremely challenging both due to high computational (memory and runtime) requirements as well as the biological complexity (genetic diversity and mobile genetic elements) [274]. Alternatively, reference databases are used for functional profiling or structural variation detection across multiple samples [271, 275, 276]. In contrast with the co-assembly approaches, the reference-based approaches are much faster and scale to large number of metagenomic samples; however, they are limited by the reference biases and can only facilitate the detection and analysis of sequences present in the reference databases. Thus, both co-assembly and reference-based approaches preclude a thorough whole-metagenome-scale microbiome analysis of large numbers of microbial samples.

Locality-sensitive hashing approaches, like the MinHash technique, have gained popularity for the study of metagenomic samples. Prominent methods include `Mash` and `Mash Screen` [260, 277]. These approaches rely on alignment-free comparisons

of whole metagenomic sequences and facilitate a rapid analysis of metagenomic samples. The resemblance or containment of genomic sequences is estimated using a small fraction of the total k-mers: both, `Mash` and `Mash Screen`, employ a 'bottom-sketch' strategy, whereby each genomic sample is represented via a *'sketch'* of k-mers—all k-mers from a genome are passed through a hash function and the genome is then represented by its sketch which comprises of the smallest hash values. Subsequent comparisons between the genomic samples are made exclusively using their respective sketches. However, these MinHash techniques have, so far, been limited to just the pairwise comparisons of metagenomic samples or for scanning a metagenomic sample across curated reference genomes. An approach like `Mash` can only estimate the similarity between two metagenomic samples, but cannot enable the detection of specific genomic regions shared between the samples. Additionally, if used on individual metagenomic assemblies, the k-mer sketches may not adequately represent all contigs, thereby hampering an all-encompassing comparison between the samples.

Here, we describe a novel approach, `SIMILE`: a fast and scalable tool that discovers the genomic regions with high sequence similarities shared between several assembled metagenomes. `SIMILE` can be parallelized over multiple threads and uses disk-based storage, enabling it to scale to hundreds of metagenomic assemblies. `SIMILE` makes two key contributions. First, it uses a fast and scalable alignment-free MinHash technique while ensuring a near-uniform sampling of k-mers throughout each metagenomic assembly. This enables each contig from the metagenomic as-

semblies to be represented in the k-mer sketch and available for comparisons. The second key contribution is that `SIMILE` provides similarity estimates and facilitates the identification of similar contigs across the metagenomic assemblies.

6.2 Methods

6.2.1 Overview, notation, and definitions

A high-level overview of `SIMILE` is shown in Fig. 6.1. `SIMILE` takes as input a set of metagenomic assemblies $\mathbf{M} = \{M_1, \ldots, M_{|\mathbf{M}|}\}$, where $|\mathbf{M}|$ denotes the total number of input assembled metagenomes. An assembled metagenome M_i corresponds to the metagenomic assembly of the sample i using its whole shotgun metagenomic sequenced reads, i.e. $M_i = \{C_{i,1}, C_{i,2} \ldots C_{i,m_i}\}$, where each $C_{i,j}$ is a contig in the assembled metagenome M_i. `SIMILE` uses an alignment-free MinHash approach to sample the k-mers—each contig $C_{i,j}$ is partitioned into bins of a fixed sized (Ω) followed by an unbiased sampling of k-mers within each bin at a user-defined k-mer sampling frequency (η), providing a near-uniform sampling of k-mers from each contig in every assembled metagenome under study. A sampled k-mer is considered to be a 'shared k-mer' if it has been sampled from at least $\epsilon \times |\mathbf{M}|$ assembled metagenomes. A k-mer-bin mapping is maintained for all sampled k-mers for each assembled metagenome and enables the determination of the shared k-mers. A contig $C_{i,j}$ is deemed a 'shared contig' if a certain minimum number of k-mers

sampled from this contig are shared k-mers; the threshold for the minimum number of shared k-mers required is determined as a function of the contig length and user choice: either a minimum length, L, of the contig should be shared or at least a proportion, ρ, of the contig length should be shared. The k-mer-bin mapping and the shared k-mers together enable the identification of shared contigs within each assembled metagenome and the similarity between contigs (i.e., contig resemblance) across the assembled metagenomes).

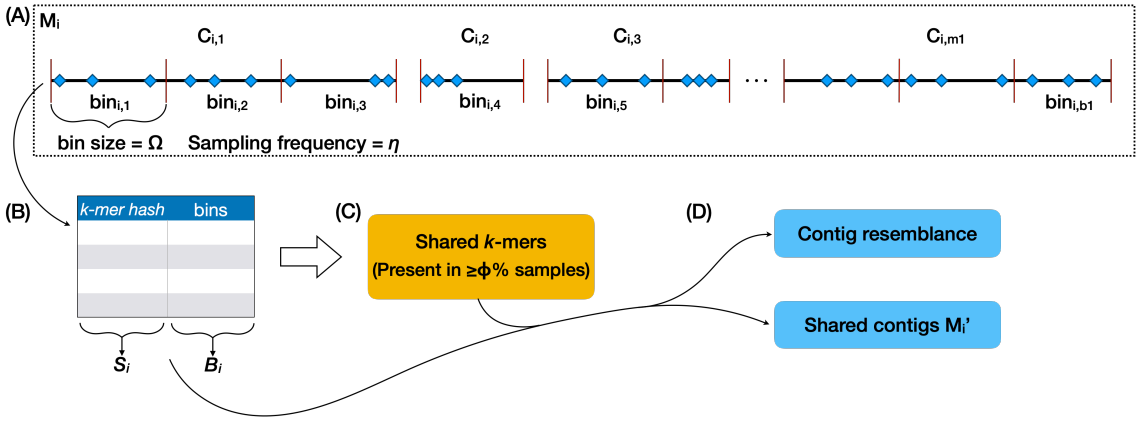


Figure 6.1: Overview of `SIMILE`. (A) Contigs ($C_{i,j}$) from each metagenomic assembly (M_i) are first binned using a fixed bin size (Ω), shown using vertical lines. A sampling frequency (η) determines the number of k-mers sampled from each bin; the sampled k-mers are shown by diamonds. (B) A mapping between the sampled k-mers and their corresponding bins is maintained for each assembly. (C) Using the sampled k-mers from all metagenomic assemblies, we identify the set of *'Shared k-mers'* that are identified in at least a certain proportion (ϵ) of the samples. (D) The shared k-mers and the individual assembly specific k-mer-bin mappings are used to determine the *'shared contigs'* from each metagenomic assembly and the resemblance between contigs from different metagenomic assemblies. The shared contigs (M_i') for each metagenomic assembly (M_i) and the contig resemblances constitute the output from `SIMILE`.

6.2.2 K-mer sampling and binning

Contrary to `Mash` where the MinHash sketching is performed using all k-mers from a metagenomic sample, `SIMILE` relies on partitioning a metagenomic sample uniformly into *bins* and computing the MinHash sketches within each *bin*. A fixed bin size, Ω (user-defined, default: 500 nt), is used to partition each assembled metagenome $M_i \in \mathbf{M}$. Within each bin, `SIMILE` determines the constituent k-mers (k is user-defined, default: 17). `SIMILE` uses canonical k-mers and each k-mer is hashed into a 64-bit value. For the nucleotide to numeric conversion, we employ the same strategy as previously used in `PRAWNS` [154]. A sampling frequency, η (user-defined, default: 5%), is used to determine the number of k-mers within each bin to be retained with the sketch. Thus, each bin contributes to $\eta\Omega$ k-mers identified via bottom-sketching, i.e. the k-mers with the smallest hash values from each bin. For an assembled metagenome of length n, the entire collection of sampled k-mers can be computed in $\mathcal{O}(n \log(\eta\Omega))$. For each metagenomic assembly, a mapping is maintained between the sampled k-mers and the bins where the k-mers were identified (Fig. 6.1).

The k-mers from each metagenomic assembly $M_i \in \mathbf{M}$ are sampled and binned independently of other assemblies. If multiple cores are available, this process can be easily parallelized. The collection of all k-mer hashes (S_i from Fig. 6.1) from a metagenomic assembly (M_i) constitute the effective k-mer sketch for the corresponding metagenomic assembly. Observe that a k-mer may be represented multiple

times within S_i if it is sampled from multiple bins in the metagenome.

The bin size (Ω) and the sampling frequency (η) together ascertain the extent of k-mers sampled and the representation for the genomic regions via k-mer sketches. A higher value for η would facilitate more k-mers to be compared, but would result in higher number of comparisons and, hence, increased run-time. A smaller Ω means smaller bins and forcing k-mers to be sampled more evenly across the contigs; however, it also increases the memory footprint with more bins generated and to be maintained in the k-mer-bin mapping.

6.2.3 Shared contigs extraction

Once the k-mer-bin mappings are obtained for each metagenome $M_i \in \mathbf{M}$, we determine the set of *'shared k-mers'*. `SIMILE` only retains the collection of k-mers which were shared between $\geq \epsilon \times |\mathbf{M}|$ metagenomes, where ϵ is a user defined threshold (default: 0.20); the shared k-mers, thus identified, are represented by $\mathcal{S}$. If $\bar{L}$ be mean total length of each metagenome, $\mathcal{S}$ can be efficiently computed in $\mathcal{O}(|\mathbf{M}|\bar{L}\log\bar{L})$. This, in turn, facilitates the identification of k-mers shared between several metagenomes: for a metagenome $M_i \in \mathbf{M}$, the sampled k-mers shared with multiple metagenomes are given by $S_i' = S_i \cap \mathcal{S}$.

The shared contigs from a metagenome M_i can be estimated using the shared sampled k-mers S_i'. `SIMILE` provides two choices for the determination of the shared

contigs. First, a contig is deemed shared contig if at least a certain minimum length L (user-defined, default: 1000 nt) of the contig is shared with contigs from other metagenomes. Second, a contig is deemed shared contig if at least a certain minimum proportion of the contig length, ρ, (user-defined, default: 40%) corresponds to the genomic region shared across metagenomes. In both the approaches, for each contig, `SIMILE` estimates the minimum number of the sampled k-mers that need to be in S'_i. For instance, for a contig $C_{i,j}$ of length $l_{i,j}$ from the metagenome M_i, $|s_{i,j}| = \eta\Omega \times \lceil l_{i,j}/\Omega \rceil$ k-mers would be sampled, where $s_{i,j}$ denote the k-mers sampled from the contig $C_{i,j}$. If shared contigs are to be identified via the shared minimum proportion (`--shared_contigs_by_proportion`) option, $\rho \times |s_{i,j}|$ or more k-mers from $s_{i,j}$ should be in S'_i. In case of locating shared contigs via minimum shared length (`--shared_contigs_by_length`), $|s_{i,j}| = \eta\Omega \times \lceil L/\Omega \rceil$. Additionally, `SIMILE` incorporates a tolerance parameter, τ, to accommodate mismatches between the shared genomic regions across different metagenomes. With both the approaches for extracting the shared contigs, `SIMILE` estimates the shared metagenome M'_i from each input metagenome $M_i \in \mathbf{M}$ such that if a contig $M'_i \in M_i$. An output file comprising of FASTA sequences is generated for the shared metagenome identified from each input metagenome.

6.2.4 Contig resemblance estimation using sequence divergence

As `SIMILE` uses an alignment-free approach, it adjudges the contig similarity using the Jaccard estimate j—a MinHash approximation for the Jaccard index, previously used in `Mash` and `Mash Screen`. If k_A and k_B are the sampled k-mers from the contigs A and B, respectively, then their Jaccard estimate $j_{(A,B)} = \frac{k_A \cap k_B}{k_A \cup k_B}$. The Jaccard estimate provides an unbiased approximation of the true Jaccard index with an error bound of $\mathcal{O}\big(1/\sqrt{\eta\Omega}\big)$.

We model the k-mer survival as a Poisson process, previously noted by Fan *et al.* [278]: if d be the probability of a single substitution, the probability of observing k consecutive nucleotides without any substitutions is e^{-kd}. With the assumption of k-mers being independent of each other, if w k-mers are conserved from a total set of z k-mers in a contig sequence, we get $e^{-kd} = \frac{w}{z}$, thus, yielding $d = -\frac{1}{k} \ln \frac{w}{z}$. Comparing two contigs from two separate metagenomic assemblies presents a set of challenges as these contigs could exhibit similarities in a variety of forms: a contig could be 'contained' within another contig, two contigs may have some overlap (shared genomic regions on either of the contig ends), or a section of the regions within the contigs may be similar or identical. To accommodate such different variations in the similarity between the contigs, we modify the approach employed in `Mash`: we frame the Jaccard estimate $j = \frac{w}{2n-w}$, where n is the length of the shorter contig. The fraction of k-mers shared between two contigs can, thus, be formulated as $\frac{w}{n} = \frac{2j}{j+1}$. The divergence between two contigs is, thus, given as

$d = -\frac{1}{k} \ln \frac{2j}{j+1}$. Note that the divergence formulation is similar to the `Mash` distance computation; however, `SIMILE` uses the Jaccard estimates to compare contigs and not entire metagenomes.

The contig similarity output is presented as a comma-separated file where each row is a 5-tuple denoting the divergence between a pair of contigs. If two contigs, $C_{i,j}$ and $C_{l,m}$, have a divergence d, the corresponding row in the output file would be $\langle i, j, l, m, d \rangle$. This contig similarity estimate can be further used to construct a graph representation, where vertices represent the metagenomic contigs and an edge $e(u, v)$ link the contigs u and v if the corresponding contigs have some similarity. The edges could be weighted edges, where edge weights denote the estimated contig divergences. Using this graph representation, we can easily identify the clusters of contigs that have some shared genomic regions. `SIMILE` provides the auxiliary files to process its primary output and extract such clusters of contigs.

6.2.5 Implementation

`SIMILE` is an open-source code (https://github.com/KiranJavkar/SIMILE.git) available under the GPLv3 license. It is implemented in C++ and Python3 and works on Unix-like operating systems. The default parameters are calibrated for analyzed metagenomes (see Results). The detailed documentation is available along with its source code.

`SIMILE` is designed to work using limited main memory (RAM usage) with the availability of disk usage and supports parallelization. Each module can be executed in parallel over *tc* cores (user-defined, default: 8), and the provision to access disk storage enables `SIMILE` to process several number of metagenomes with large total length of the metagenomes. If N be the cumulative total length of the assembled metagenomes and L_{max} be the maximum total length among the input metagenomes, then `SIMILE`'s run-time complexity is $\mathcal{O}\big(N \log(\eta\Omega L_{max})\big)$. If the total length of each metagenome is roughly similar ($N \approx |\mathbf{M}|\bar{L}$), the run-time is linear in the number of input metagenomes.

6.2.6 Analysis

6.2.6.1 Assembly

Assembled metagenomes for human stool samples were downloaded from Human Microbiome Project (HMP) website [268]. Sequencing reads for additionally analyzed whole genomes and metagenomes were downloaded from NCBI. The genomes for pure-culture isolates were assembled with SPAdes (v3.13.0, with `-careful -cov-cutoff auto`) [144]. The metagenomic samples were assembled using Megahit (v1.1.3) [249]. In order to ensure better confidence in the contigs and the subsequent analysis from the assembled genomes and metagenomes, only the contigs with at least 500 bp in length were retained. The NCBI identifiers for all samples used are

available in Supplementary Table 1.

6.2.6.2 Ground truth

Due to the absence of an actual ground truth, we compared the results from `SIMILE` to those obtained by aggregating the all-vs-all pairwise BLAST (blastn v2.8.1) comparisons [156]. The genomic regions were deemed shared between the corresponding pairs of assembled metagenomes if the alignments had a sequence identity of at least 95%; the respective genomic regions are referred to as 'pairwise shared regions'. To gauge the genomic regions shared between at least $\epsilon \times |\mathbf{M}|$ assembled metagenomes, we create a pile-up for the pairwise shared regions identified for each M_i. Using this pile-up, we check for the genomic regions that are shared with $\geq \epsilon \times |\mathbf{M}|$ metagenomes; these constitute the shared genomic regions. Next, we compute the effective shared length for each contig and determine if the contig is to be deemed a 'true shared contig', depending on the user choice for the type of shared contigs to be extracted, i.e., by length or by proportion, and the input thresholds (L or ρ).

6.3 Results

6.3.1 Methods compared

As of the time of writing this manuscript, we are unaware of any other tool that can compare several assembled metagenomes together and identify similar genomic regions shared between the metagenomes. Therefore, we benchmarked `SIMILE` against the all-vs-all pairwise comparisons with BLAST; the BLAST comparisons were also used to determine the ground truth, as described in the Methods section. As BLAST yields highest accuracy among the available tools for pairwise sequence comparisons and alignment, the benchmarking comparisons were limited to BLAST. All tools were run on a 64 core Xeon E5-2680 server running at 2.70 GHz and a total of 256 GB of RAM.

6.3.2 Performance

6.3.2.1 Scalability

We benchmarked the performance of `SIMILE` using the assembled metagenomes for 208 stool samples from the Human Microbiome Project, hereafter referred to as the HMP stool dataset. The HMP stool dataset was randomly sampled without replacement to create smaller datasets of 25, 50, 75, 100, 125, 150, 175, and 200

assembled metagenomes. SIMILE was run with default parameters (k=17) and the maximum memory usage was limited to 36GB. Both, SIMILE and BLAST were executed using 8 cores. Figure 6.2 shows the run-time performance and scalability of SIMILE to extract shared contigs (by minimum contig proportion), contrasted against the all-vs-all pairwise BLAST comparisons. SIMILE scaled to handle and extract shared contigs from the entire HMP stool dataset with its run-time linear in the number of input metagenomes. In comparison, the all-vs-all pairwise BLAST comparisons had a run-time quadratic in the number of input metagenomes and did not scale well to higher input metagenome counts. The run-time performance was comparable for both the choices of extracting shared contigs—by minimum shared length as well as by minimum contig proportion; in both the cases, SIMILE extracted shared contigs from the complete HMP stool dataset (n=208) in less than 4 hours and under 21.4 GB of peak memory usage (Supplementary Figure 1).

6.3.2.2 Accuracy

Figure 6.3 shows the impact of the choice of user-defined parameters—k-mer length, sampling frequency (η), and bin size (Ω)—on the accuracy of shared contigs extracted using SIMILE; all corresponding values for the precision, recall, and run-time are tabulated in Supplementary Table 2. Here, we benchmarked the performance on 25 assembled metagenomes from the HMP stool dataset and the accuracies were calculated with respect to the ground truth computed using BLAST

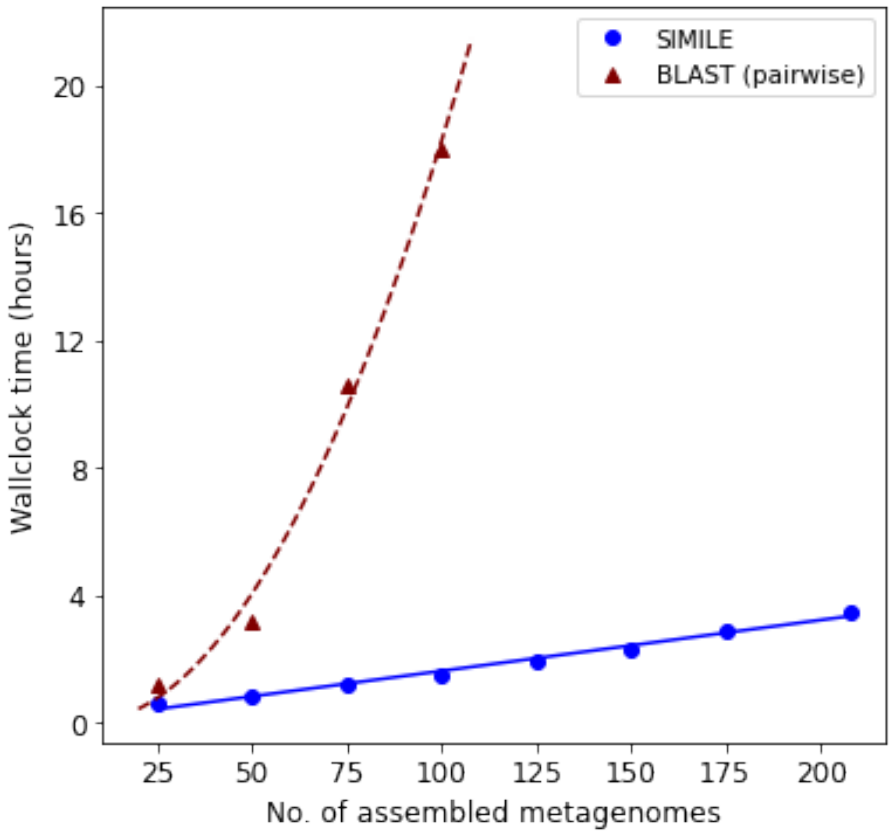


Figure 6.2: Scalability performance using the HMP stool dataset. The plot shows the runtime performance of `SIMILE`, contrasted against the all-vs-all pairwise BLAST comparisons, using 8 cores on the HMP stool dataset comprising 208 assembled metagenomes.

comparisons (see Methods). For a k-mer length of 15, `SIMILE` performed at >99% recall, but with a very low (approx. 48%) precision, for a bin size of 500 bp and 1000 bp. For k-mer lengths from 17 to 21, the accuracies were comparable: for a bin size of 500 bp, as the k-mer length increased from 17 to 21, the precision increased from 80% to 88% at the cost of recall dropping from 95% to 88%. The bin size (Ω) inversely impacted the recall performance of the shared contigs: a smaller bin size (500 bp) enforced more uniform sampling of k-mers than a larger bin size (1000 bp). At k=17 and η=5%, the recall was 95% and 94% for a bin size of 500 bp and 1000 bp respectively. In contrast, at k=21 and η=5%, the recall was 88% and 85% for the

respective bin sizes, suggesting a higher k-mer length would have a larger difference in recall arising due to changes in bin size. The sampling frequency (η) determines the number of k-mers to be sampled and, hence, directly influenced the accuracy of shared contig extraction. For k=17 and Ω=500bp, at η=1%, the precision and recall was 76% and 89% respectively, whereas at η=5%, the values for the same are 80% and 95% respectively. However, the performance gains are marginal when the sampling frequency was increased further: as η was increased from 10% to 20%, the recall increased from 96% to 97%, while the precision stayed at 80%. The increase in sampling frequency also increased the run-time: for η=1%, the `SIMILE` run took 22 minutes, whereas for η=20%, the corresponding run took 32 minutes.

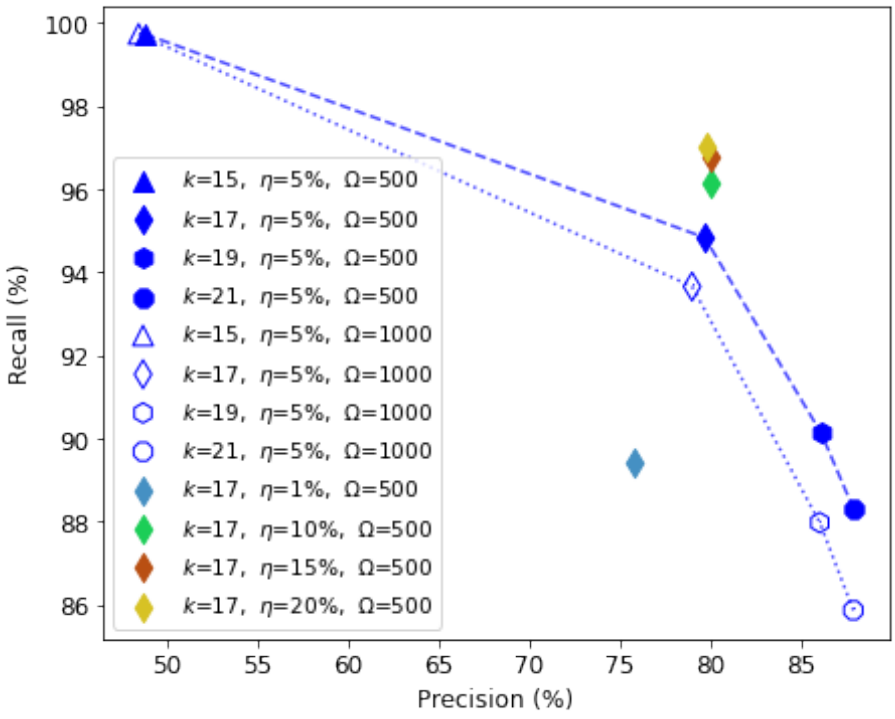


Figure 6.3: Precision-Recall performance of `SIMILE` by varying its parameters—k-mer length, sampling frequency (η), and bin size (Ω)—on 25 assembled metagenomes from the HMP stool dataset.

6.3.3 Applications

6.3.3.1 Genomic regions shared between metagenomic samples

To demonstrate the extent of metagenomic information we lose by limiting the whole shotgun metagenomic analysis to just the contigs assigned to known taxa, we checked for the taxa detected within the input assembled metagenomes and in the shared contigs from `SIMILE`. The taxa were identified using KRAKEN [205]. Figure 6.4 shows the median proportions of the top 25 taxa, by total contig length and total number of contigs, as identified in the input assembled metagenomes from the entire HMP stool dataset (n=208) and the corresponding shared contigs (by minimum contig proportion) extracted using `SIMILE` with default parameters (k=17). In Figure 6.4, the taxa statistics for the input assembled metagenomes are denoted as 'Initial', while those for the shared contigs are denoted as 'SIMILE'. The input assembled metagenomes had a mean total sequence length of 136 Mbp and 58,317 mean total number of contigs per assembled metagenome. With the shared contig extraction using `SIMILE`, we obtained a mean total shared sequence length of 86 Mbp and 23,689 shared contig count per metagenome. For the input assembled metagenomes, the unclassified sequence data corresponded to 26% of the median total sequence length and 43% of the median total number of contigs per metagenome. After the extraction of shared contigs, the unclassified sequence data attributed to 44% of the median total shared sequence length and 66% of median

shared contigs per metagenome. In other words, out of the genomic regions shared between $\geq 42/208$ (20%) HMP stool samples, on an average, 44% genome sequence (66% by contig counts) cannot be classified into any of the known taxa. The genomic regions associated with unclassified taxa often get excluded from analyses in taxa presence-absence driven metagenomic studies [279, 280].

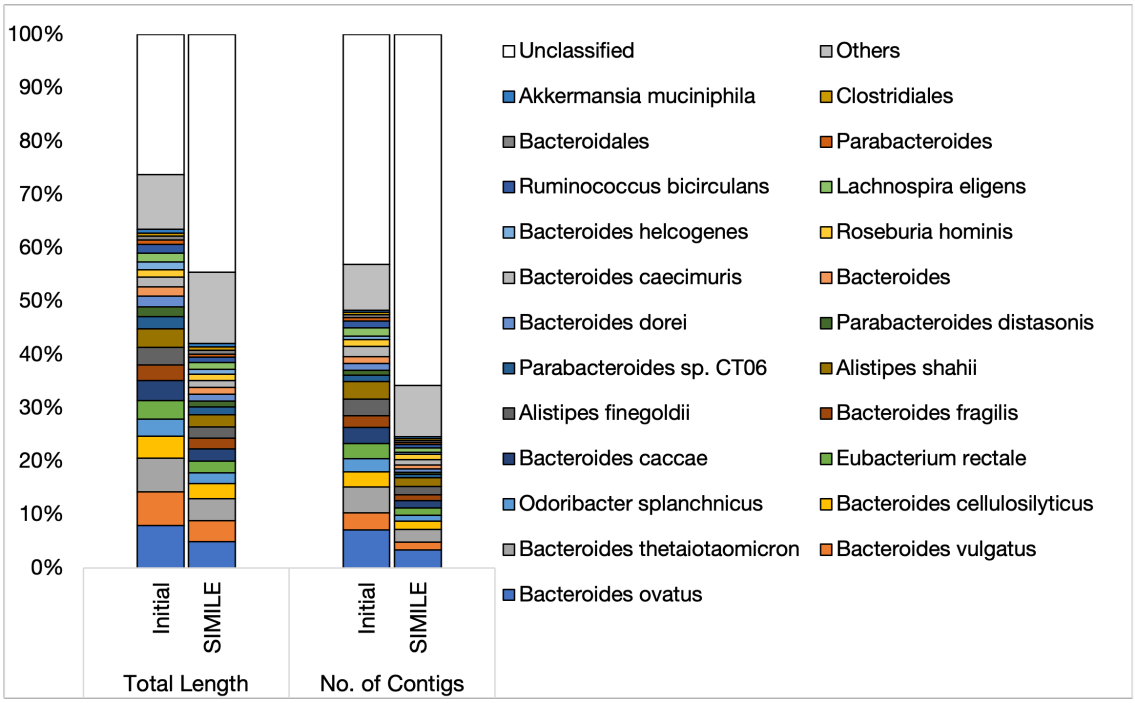


Figure 6.4: Median proportion of contigs, by total length and counts, in 208 HMP stool samples categorized by the corresponding taxa. The total length and contig counts of top 25 taxa are contrasted from the initial assembled metagenomes for the HMP stool samples (Initial) against the shared contigs extracted using `SIMILE` (SIMILE).

`SIMILE` also facilitates the estimation of contig resemblance and assessing the similarity between contigs from the input assembled metagenomes. `SIMILE` provides a similarity estimate between pairs of contigs from the input metagenomes; this similarity estimate can be used to cluster and extract the contigs containing putatively similar genomic regions. Figure 6.5 shows an example of contigs associated

with genomic regions shared between 12 input metagenomes (Supplementary Table 3). The highlighted regions correspond to the genomic regions with ≥95% sequence identity. The longest contig in this contig cluster was 227 kbp in length while the contig alignment spanned 139 kbp. Despite these genomic regions being shared between several metagenomic samples at a high sequence similarity, not much was known about the corresponding genomic regions. NCBI web BLAST comparisons suggested these shared contigs to have weak similarity to *Dialister* species (≤3% query coverage).

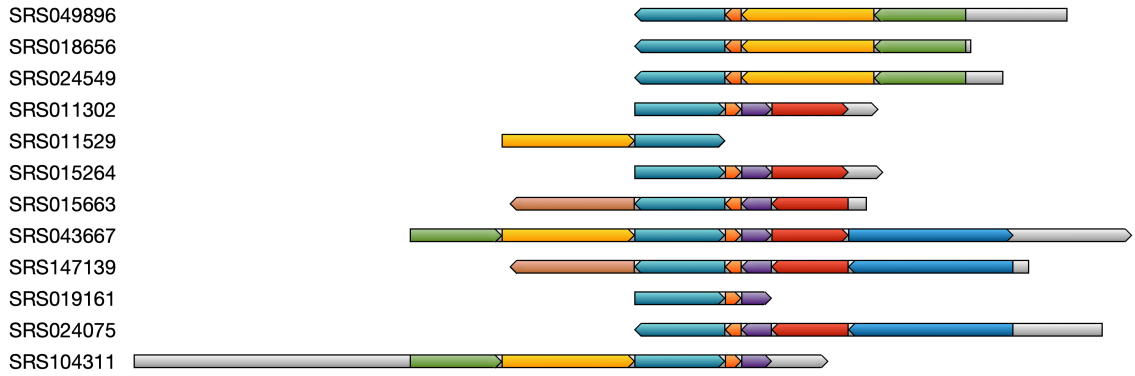


Figure 6.5: Cluster of similar contigs identified using `SIMILE`'s similarity estimate. The highlighted regions denote the regions with BLAST matches (≥95% sequence identity) between the corresponding contigs. The contig alignment spanned 139 kbp region and the longest contig had a length of 227 kbp. NCBI web BLAST for these contigs showed weak similarity with *Dialister* species (≤3% query coverage)

6.3.3.2 Case-control analysis

For a proof-of-concept case-control analysis, we used the 55 whole shotgun metagenomic gut samples from Vangay *et al.* [95]. These samples correspond to 55 females, included individual from Hmong, Karen, and the US. The study seeked to

uncover the differences in the gut microbiome as individuals migrated to the US; the samples from US-based individuals (European ancestry) formed the controls in the analysis.

We constructed assembled metagenomes for each of these gut samples and ran `SIMILE` (default parameters) to extract the shared contigs; `SIMILE` was run on 8 cores each of 2.30 GHz Xeon E5-2650 processor with 36 GB maximum RAM limit and the execution completed in 2h 15m. The original study had performed taxa presence-absence analysis and inferred that the gut samples from US-based individuals had varied *Bacteroides* strains while those from Hmong had varied *Prevotella* strains. The shared contigs identified using `SIMILE` supported a similar observation—the Kraken taxa located *Bacteroides* strains in the samples from US-based individuals and *Prevotella sp.* in those from Hmong ones. Additionally, using the contig resemblance estimate, a contig cluster corresponding to *B. vulgatus* was identified in metagenomes for the controls (Supplementary Table 3). In the samples from Hmong, the contig resemblance estimation enabled the detection of a contig cluster that corresponded to *Lachnospiraceae bacterium*, putatively *Ruminococcus gnavus*. The role of *Lachnospiraceae* bacteria, including *R. gnavus*, has not well understood in human health—they have been detected in infant gut as well as reported to be associated with diseases having inflammatory conditions, like IBD [281, 282, 283] (Supplementary Table 3).

6.3.3.3 Contamination screening

The shared contig extraction with `SIMILE` can also be used for a rapid contamination screening and filtering. As a proof of concept, we created a mock whole genome dataset of 120 clinical bacterial isolates: 100 *Salmonella enterica* and 20 *Escherichia coli*. The mock dataset represented a genome collection 120 samples comprising *S. enterica* genomes 'contaminated' with *E. coli* genomes.

Figure 6.6 shows the genome length distributions of initial mock dataset ('Initial') and the shared contigs by minimum contig proportion extracted using `SIMILE` ('SIMILE') with default parameters. `SIMILE` was executed on this mock dataset (n=120) with the intention of removing most of the contigs from 'contaminating samples' (i.e., the 20 *E. coli* genomes) while retaining most of the contigs from the 'non-contaminant samples' (i.e., the 100 *S. enterica* genomes). In other words, the contigs from *S. enterica* genomes would be deemed shared contigs and retained by `SIMILE`, whereas those from *E. coli* would be discared as they would be present in fewer samples than required by the thresholds used. The median total lengths of *S. enterica* and *E. coli* genomes were 4.75 Mbp and 5.1 Mbp respectively, resulting in the combined mock dataset to have a median total length of 4.80 Mbp. On extraction of shared contigs, the combined dataset had a median total length of 3.91 Mbp—shared contigs from *S. enterica* had a median total length of 4.69 Mbp whereas those from *E. coli* had a total median length of just 5.9 kbp, i.e. retaining most of the *S. enterica* contigs while filtering out those from *E. coli*. `SIMILE` was

run on 8 cores each of 2.30 GHz Xeon E5-2650 processor with 36 GB maximum RAM limit and the execution completed in 37 minutes.

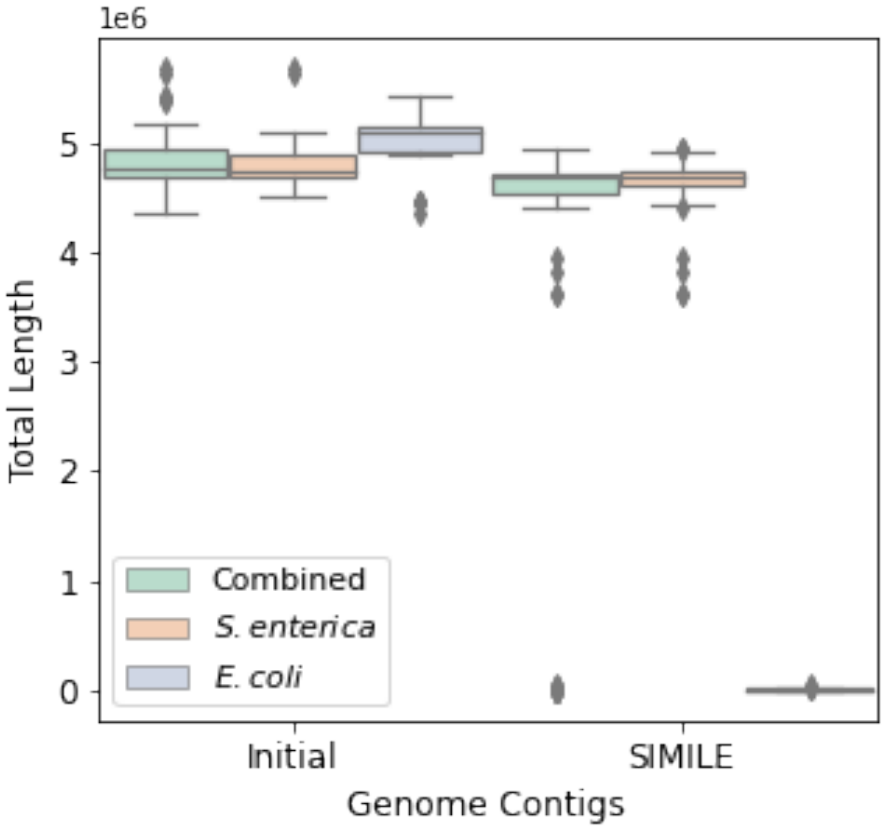


Figure 6.6: Total genome length distributions of a whole-genome dataset ($n = 120$) of *Salmonella enterica* genomes ($n = 100$) 'contaminated' with *Escherichia coli* genomes ($n = 20$). The total genome length for the *S. enterica* and *E. coli* isolates were comparable in the initial assemblies. The extraction of shared genomic regions by `SIMILE` from this dataset maintained the total genome length of *S. enterica* isolates to be similar to their initial total length, whereas most of the contigs from *E. coli* isolates were filtered out.

6.4 Discussion

We developed `SIMILE` to fill a gap in the current tool-kit available to scientists for comparing several whole shotgun metagenomic samples. Co-assembly

approaches for multiple metagenomic sample comparisons do not scale beyond a tens of samples, whereas the reference-based approaches are inherently limited by the reference biases and do not support the assessment of novel or lesser studied microbes. Studies relying on taxa presence-absence discard the assessment of genomic variations which may be associated or important for microbiome analysis, particularly in case-control studies. With `SIMILE`, we present a scalable alternative for large-scale whole shotgun metagenomic analysis that can scale to hundreds of metagenomic samples and can identify the genomic regions shared within multiple samples, facilitating the identification of similar genomic regions and enabling the detection of relevant genomic regions for a thorough downstream assessment. Similar to population genomics where a collection of closely related whole genomes are analyzed for their genomic similarities or differences, the detection of shared genomic regions from several metagenomic samples can motivate the field of population metagenomics for metagenome-wide association studies analyzing the similarities or differences between multiple metagenomic samples [21, 154].

`SIMILE` takes as input a collection of sample-specific assembled metagenomes; construction of such assembled metagenomes are far less computationally expensive than a co-assembly approach as they can be constructed in parallel and with much less compute requirements. The alignment-free implementation enables `SIMILE` to scale to increasing numbers of metagenomic samples. When applied to real biological datasets, `SIMILE` could identify the genomic regions shared in multiple samples, without the need for any reference. As demonstrated in the results, even in rel-

atively better studied microbial communities, like the human gut microbiome, a substantial proportion of the microbiome comprises of unknown or uncharacterized taxa. The shared contig extraction and contig resemblance estimation would facilitate a better identification and understanding of the corresponding genomic regions. This would provide important insights in both verifying the taxa associations reported in previous studies as well as the overall understanding of complex microbial communities.

Identification of metagenome-assembled-genomes (MAGs) from microbial communities or metagenomic samples would highly benefit from the detection of genomic regions shared in multiple samples. The increasing availability of whole shotgun metagenomic samples and the improvements in sequencing technologies to support metagenomics have incentivized the extraction of MAGs; recent studies have used MAGs to uncover the genomic factors associated with geospatial distribution of strains of microbial species that are not well characterized [283]. The detection of shared genomic regions facilitate better identification of MAGs and can motivate the development of tools better equipped to analyze MAGs and overall population metagenomics.

Chapter 7: Conclusion

In this dissertation, we described several challenges in large-scale whole-genome analyses and presented novel methods for addressing them. We believe our tools and algorithms will aid in extracting meaningful information from large genomic datasets. We have demonstrated the practicality of our analyses and tools on real genomic and metagenomic datasets.

Antimicrobial resistance continues to increase worldwide with acute impacts on public health in developing as well as developed countries [25, 26, 27]. It is only a matter of time until the currently effective antibiotics become ineffective in treating pathogenic infections [28]. In order to develop novel antibiotics effective in treatment of these bacterial infections, it is important to understand the evolution of the pathogenic isolates resisting the actions of existing drugs. Our works on genomic factors associated with antimicrobial resistance highlight the intra- and extra-chromosomal genomic variations—apart from the known AMR factors—that may influence the AMR phenotype [20, 21]. Many genomic variants associated with AMR have been observed on plasmids and acquired via horizontal gene transfer events. With our analyses, we underscore instances where chromosomal factors

determine whether an isolate can accommodate such plasmids, and cases where typical AMR-associated genomic variants are present in the susceptible isolates rather than the resistant ones [20, 21]. Our works show that a large-scale analysis can highlight instances for targeted high-resolution comparisons and discover putatively novel genomic features differentiating the phenotypes. Identification of such genomic features will also motivate the characterization of relatively understudied genomic regions, including the vast number of hypothetical proteins, intergenic regions, replicons, insertion sequences, and promoter region changes. A thorough understanding of the genomic determinants of AMR is extremely important in understanding the efficacy of antibiotics and mitigating the infectious diseases.

As sequencing technologies continue to improve and become more accessible, genome sequencing of pathogenic microbes should be incentivized worldwide and adapted in extensively various settings—particularly in public health and infectious disease monitoring—as it would facilitate better monitoring and regulating of the spread of AMR factors and pathogens in food and healthcare settings. [86]. Efficient use of genomics for pathogen detection and public health programs benefits from accurate and timely pathogen genome reconstruction. Our work on quasimetagenomics demonstrated that pathogen genome reconstruction and analysis can be expedited; a combination of different sequencing technologies help in obtaining a highly accurate assembled genome without a substantial fiscal cost [99]. Rapid identification of pathogens is vital in not just containing the spread of infectious diseases, but also identifying specific genomic factors attributed to pathogenicity,

virulence, or AMR, enabling us to implement targeted disease therapeutics.

The astronomical increase in availability of whole-genome sequenced microbes has facilitated insights into microbial genome biology. Although a disproportionate number of genomes are available for a select few taxa, it has presented novel challenges for large-scale bioinformatics analysis tools. Currently, the choice of bioinformatics tools is based on whether the genomes under study have an open or closed pan-genome and the computational resources available. As more genomes are sequenced, the concept of an open or closed pan-genome needs to undergo careful scrutiny. For example, like in the case of our work on *A. baumannii*, specific clades of isolates may have a closed pan-genome, but the species pan-genome may be open [21]. Newer bioinformatics tools should support the detection of such genomic similarities and differences, and aid in associated downstream analyses.

Microbial genomics continues to suffer from the 'large p, small n' problem, i.e., the number of genome sequences with relevant metadata is relatively small compared to the number of genomic features that may potentially explain the observed phenotypes. With `PRAWNS`, we have presented an alternative to scale down the genomic features by an order of magnitude compared to that obtained using just k-mers or exact matches [154]. Although these features can support better assessment of genotype-phenotype associations, more needs to be done to condense the feature counts while ensuring that these genomic features do not over-generalize, hiding the important and sensitive structural variations that differentiate the genomic regions with high sequence similarity.

Many tools developed to support population genomics or large-scale genome analyses rely on the clustering or aggregation of sequences into a single entity (for example, a cluster center sequence representing all gene sequences in that cluster) which then forms the basis for subsequent comparisons [49, 284]. Often, these tools employ a single greedy strategy for aggregating the corresponding sequences. However, functional characteristics and modifications cannot be obviously distinguished and associated with all genomic variations. For instance, two AMR gene sequences could just be one SNP apart and may exert different degrees of resistance, whereas other genes may have much lower sequence similarity yet be functionally identical [51]. Alternatively, genes could be shared between multiple organisms but have intragenic regions that support variable degrees of taxonomic classification [96]. Newer tools should account for such dynamic variability between different genomic regions to support better extraction of biologically meaningful and relevant information.

Genotype-phenotype correlation studies have relied on a few statistical frameworks (metric multi-dimensionality scaling, linear mixed models, etc.) to estimate the effect of population structure and gauge the genomic variations putatively explaining the phenotypes. Owing to the bottlenecks of the underlying models and the 'large p, small n' situation, different tools use various approaches and heuristics to limit the number of features to be analyzed [65, 66]. Furthermore, several approaches assume feature independence to reduce the computational complexity—the statistical significance of each genomic feature is, thus, assessed independent of other features, discarding the influence of genomic context. As `PRAWNS` gener-

ates a concise collection of features, it presents an opportunity to formulate better statistical tests where the features can be evaluated together, while accounting for their genomic context and interactions between the corresponding regions. These features also allow for the construction of phylogenetic trees and the visualization of variability across the isolates under study. Such visualizations would facilitate targeted and robust assessments rather than enforcing generalized population structure estimation and significance analysis.

As more and more sequenced genomes continue to become available, population genomics promise large-scale longitudinal analyses as well [285]. Longitudinal population genomics can facilitate the exploration of evolution of the isolates, emergence of novel clades and structural variations, and the study of their corresponding phenotypes. Existing tools for population genomics, such as `PRAWNS`, require longitudinal data to be processed in the downstream analyses as an auxiliary metadata associated with the genome sequences. The availability of abundant and accurate (spatio-)temporal datasets could motivate the development of tools and approaches where the metadata is incorporated within the comparative analysis. In addition to identification of clades of genomically similar isolates, longitudinal population genomics could support the visualization and analysis of contemporary isolates; this would support the evolutionary analysis of genomes undergoing phenotypic variations and allow for the variants to be monitored over time [286].

The improvements in metagenomic sequencing and assembly approaches have uncovered the opportunities to extract metagenome-assembled-genomes (MAGs).

Recent studies have shown the utility of MAGs to support large-scale population genomics and identify various structural variations, including those associated with geospatial phenotypes [283]. The analysis of MAGs presents a diverse set of challenges—they are unlikely to be as complete and accurate as pure culture isolate genomes, but can provide much more cohesive genomic information about taxa, particularly those that cannot be clinically cultured, from metagenomic assemblies directly. Among the existing tools for large-scale sequence comparisons, those requiring a high genome sequence similarity would not be well-suited to compare MAGs since many of these approaches require a high average nucleotide identity (ANI) or maximal unique matches (MUMs) shared by all genomes being compared [38, 42] In contrast, the approaches catered for metagenomic comparisons may overestimate the sequence diversity between the MAGs to be compared and, hence, under-utilize the available information [260]. Large-scale analyses using MAGs would require population genomics tools using incomplete genomes with high variations. A growing impetus for such analyses would necessitate the development of new algorithms and approaches to support such comparisons.

Approaches for large-scale genomic comparisons have often been categorized broadly as alignment-based or alignment-free. The alignment-based approaches maintain the genomic context but are computationally expensive, compared to the alignment-free approaches that are extremely fast and scalable, but lose the genomic context. With our methods, we demonstrate the possibility of developing approaches that provide a middle ground. Both, `PRAWNS` and `SIMILE`, rely on k-mer

based comparisons but also maintain the genomic context of these exact matching regions [154, 266]. This allows for the scalability to genomes or metagenomes like the alignment-free approaches, while providing a functionality comparable to the alignment-based counterparts. Newer tools could be developed using such frameworks, where the initial computation could be performed using k-mers or alignment-free comparisons to get rapid scalability, followed by targeted high-resolution comparisons within the subsets of genomically similar samples.

Retrieving an accurate genome from a microbial community is often challenging and time-consuming, often limiting the efficacy of tasks like limiting the easy applicability to tasks such as pathogen detection for food-borne disease monitoring. Our work on quasimetagenomics shows that it is possible to achieve a middle ground between the laborious culture-based pure colony bacterial genomics and culture-independent metagenomics [99]. Depending on the genomic analysis to be performed, quasimetagenomics can be optimized for the amount of selective enrichment and type of sequencing—expediting the genome recovery and analysis. As sequencing technologies continue to evolve and reference databases become taxonomically comprehensive, quasimetagenomics as well as whole shotgun metagenomic sequencing could enable the assessment and understanding of different microbial communities, evolution of strains within these communities, as well as the characterization and classification of microeukaryotes.

Mobile genetic elements (MGEs), such as plasmids, phages, insertion sequences, are important genomic variations that can influence key characteristics of bacterial

isolates, including the adaptability to various habitats, gene regulation, pathogenicity and antimicrobial resistance. Understanding mobile genetic elements is an important facet in understanding microbial genome biology, particularly since they facilitate the exchange of genetic material between different isolates, even across taxa, and can influence the evolution of different strains. However, MGE analysis is quite challenging. As demonstrated in our studies, MGEs often pose challenges to the assembly algorithms. Determining the genomic context of these mobile genetic elements becomes problematic, hampering a comprehensive genome-wide analysis. Additionally, plasmids—the primary vehicles of horizontal gene transfer in bacteria—can be extremely dynamic in terms of their transmissibility and the encoded genomic regions, further increasing the complexity of characterizing such MGEs [287, 288]. With our analyses, we highlight the impact and associations of various mobile genetic elements in clinically important phenotypes and the need for their better understanding for an improved public health response [20, 21].

Analysis of mobile genetic elements (MGEs) necessitates the development of improved bioinformatic tools, approaches, and studies on several extensively sampled microbiomes which host a variety of MGEs. With the availability of diverse and improved sequencing technologies, such as Hi-C sequencing and HiFi long reads, the assembly algorithms need to be better equipped to handle the intra- and extra-chromosomal mobile genetic elements. For instance, HiFi reads can enable the assembly of entire MGEs, while Hi-C reads can facilitate the association between MGEs and chromosomes [289, 290]. Extensive sampling and large-scale studies

would facilitate the explorations of population dynamics, transmission dynamics, and the evolution of different mobile genetic elements, alongside the overall genomic variations, in both clinical and environmental settings. This would also motivate the development and maintenance of better curated reference databases and typing schemes for MGE analysis. The development of such approaches catered for MGEs would not only improve our overall understanding of microbial genome biology but also facilitate better food-safety and infectious disease monitoring networks [14].

Lastly, some of my other non-thesis works focus on the human co-operation and analysis using open datasets [291, 292]. Open datasets and publicly available sequenced data from various consortium studies have considerable steered the advancements in the field of genomics and metagenomics [268, 293]. In many other fields, open datasets with human-related data have provided important insights into human behavior and enabled the development of human-centric response programs [291, 294]. However, the analyses and inferences using open datasets have always been subject to privacy and security concerns, thereby requiring different ethical standards with a constant need for scrutiny [295]. Our work encapsulated the skepticism and doubts of people pertaining to human subject involvement in genetic/genomic research [292]; the laws for genetic research and sharing genetic data are not well-established worldwide, and the user cooperation relies on the their trust in researchers and organizations acquring their genetic/genomic data [296, 297]. As for human microbiome and metagenomics, we do not have many protocols for ethical analysis. A substantial proportion of microbial genomic data is still referred to as

the 'microbial dark matter' [298, 299]; further research would provide leads to understanding more about the associated microbiome. As genomics, metagenomics, and microbiome analyses progress, shared human microbiome data may leak personally identifiable information, as observed previously with other open datasets. Along with the developments and improvements in these fields, appropriate measures are needed that ensure no severe lapses associated with microbial research.

Bibliography

[1] Manish Boolchandani, Alaric W D'Souza, and Gautam Dantas. Sequencing-based methods and resources to study antimicrobial resistance. Nature Reviews Genetics, page 1, 2019.

[2] Jack A Gilbert, Martin J Blaser, J Gregory Caporaso, Janet K Jansson, Susan V Lynch, and Rob Knight. Current understanding of the human microbiome. Nature medicine, 24(4):392–400, 2018.

[3] Ankit Gupta, Rasna Gupta, and Ram Lakhan Singh. Microbes and environment. In Principles and applications of environmental biotechnology for a sustainable future, pages 43–84. Springer, 2017.

[4] Fabien J Cousin, Rozenn Le Guellec, Margot Schlusselhuber, Marion Dalmasso, Jean-Marie Laplace, and Marina Cretenet. Microorganisms in fermented apple beverages: current knowledge and future directions. Microorganisms, 5(3):39, 2017.

[5] Lora V Hooper and Jeffrey I Gordon. Commensal host-bacterial relationships in the gut. Science, 292(5519):1115–1118, 2001.

[6] Peyman Alavi, Margaret R Starcher, Gerhard G Thallinger, Christin Zachow, Henry Müller, and Gabriele Berg. Stenotrophomonas comparative genomics reveals genes and functions that differentiate beneficial and pathogenic bacteria. BMC genomics, 15(1):482, 2014.

[7] Leo Eberl and Peter Vandamme. Members of the genus burkholderia: good and bad guys. F1000Research, 5, 2016.

[8] Niranjan Nagarajan and Mihai Pop. Sequence assembly demystified. Nature Reviews Genetics, 14(3):157–167, 2013.

[9] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. Science, 269(5223):496–512, 1995.

[10] Gordon Luikart, Phillip R England, David Tallmon, Steve Jordan, and Pierre Taberlet. The power and promise of population genomics: from genotyping to genome typing. Nature reviews genetics, 4(12):981–994, 2003.

[11] E. Fidelma Boyd, Salvador Almagro-Moreno, and Michelle A. Parent. Genomic islands are dynamic, ancient integrative elements in bacterial evolution. Trends in Microbiology, 17(2):47–53, 2009. ISSN 0966-842X. doi: https://doi.org/10.1016/j.tim.2008.11.003. URL https://www.sciencedirect.com/science/article/pii/S0966842X09000031.

[12] Samuel K Sheppard, David S Guttman, and J Ross Fitzgerald. Population genomics of bacterial host adaptation. Nature Reviews Genetics, 19(9):549, 2018.

[13] Gregory L. Armstrong, Duncan R. MacCannell, Jill Taylor, Heather A. Carleton, Elizabeth B. Neuhaus, Richard S. Bradbury, James E. Posey, and Marta Gwinn. Pathogen genomics in public health. New England Journal of Medicine, 381(26):2569–2580, 2019. doi: 10.1056/NEJMsr1813907. URL https://doi.org/10.1056/NEJMsr1813907. PMID: 31881145.

[14] Ruth E Timme, Hugh Rand, Maria Sanchez Leon, Maria Hoffmann, Errol Strain, Marc Allard, Dwayne Roberson, and Joseph D Baugher. Genometrakr proficiency testing for foodborne pathogen surveillance: an exercise from 2015. Microbial genomics, 4(7), 2018.

[15] Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences, 102(39):13950–13955, 2005.

[16] N Stoesser, EM Batty, DW Eyre, M Morgan, DH Wyllie, C Del Ojo Elias, James R Johnson, AS Walker, TEA Peto, and DW Crook. Predicting antimicrobial susceptibilities for escherichia coli and klebsiella pneumoniae isolates using whole genomic sequence data. Journal of Antimicrobial Chemotherapy, 68(10):2234–2244, 2013.

[17] Henrik Hasman, Dhany Saputra, Thomas Sicheritz-Pontén, Ole Lund, Christina Aaby Svendsen, Niels Frimodt-Møller, and Frank M Aarestrup. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. Journal of clinical microbiology, 52(1): 139–146, 2014.

[18] J. Mitchell. Streptococcus mitis: walking the line between commensalism and pathogenesis. Molecular Oral Microbiology, 26(2):89–98, 2011. doi: https://doi.org/10.1111/j.2041-1014.2010.00601.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-1014.2010.00601.x.

[19] Salvatore Cosentino, Mette Voldby Larsen, Frank Møller Aarestrup, and Ole Lund. Pathogenfinder-distinguishing friend from foe using bacterial whole genome sequence data. PloS one, 8(10):e77302, 2013.

[20] Jay Noboru Worley, Kiran Javkar, Maria Hoffmann, Kristen Hysell, Amanda Garcia-Williams, Kaitlin Tagg, Sanjat Kanjilal, Errol Strain, Mihai Pop, Marc Allard, Louise Francois Watkins, Lynn Bry, and Peter Gilligan. Genomic drivers of multidrug-resistant shigella affecting vulnerable patient populations in the united states and abroad. mBio, 12(1):e03188–20, 2021. doi: 10.1128/mBio.03188-20. URL https://journals.asm.org/doi/abs/10.1128/mBio.03188-20.

[21] Kiran Javkar, Hugh Rand, Maria Hoffmann, Yan Luo, Saul Sarria, Nagarajan Thirunavukkarasu, Christine A. Pillai, Patrick McGann, J. Kristie Johnson, Errol Strain, and Mihai Pop. Whole-genome assessment of clinical acinetobacter baumannii isolates uncovers potentially novel factors influencing carbapenem resistance. Frontiers in Microbiology, 12:2826, 2021. ISSN 1664-302X. doi: 10.3389/fmicb.2021.714284. URL https://www.frontiersin.org/article/10.3389/fmicb.2021.714284.

[22] Michael S. Donnenberg, Tracy H. Hazen, Tamer H. Farag, Sandra Panchalingam, Martin Antonio, Anowar Hossain, Inacio Mandomando, John Benjamin Ochieng, Thandavarayan Ramamurthy, Boubou Tamboura, Anita Zaidi, Myron M. Levine, Karen Kotloff, David A. Rasko, and James P. Nataro. Bacterial factors associated with lethal outcome of enteropathogenic escherichia coli infection: Genomic case-control studies. PLOS Neglected Tropical Diseases, 9(5):1–11, 05 2015. doi: 10.1371/journal.pntd.0003791. URL https://doi.org/10.1371/journal.pntd.0003791.

[23] Abigail M. Deaven, Christina M. Ferreira, Elizabeth A. Reed, Jeremy R. Chen See, Nora A. Lee, Eduardo Almaraz, Paula C. Rios, Jacob G. Marogi, Regina Lamendella, Jie Zheng, Rebecca L. Bell, Nikki W. Shariat, and Charles M. Dozois. Salmonella genomics and population analyses reveal high inter- and intraserovar diversity in freshwater. Applied and Environmental Microbiology, 87(6):e02594–20, 2021. doi: 10.1128/AEM.02594-20. URL https://journals.asm.org/doi/abs/10.1128/AEM.02594-20.

[24] Alison Laufer Halpin, L. Clifford McDonald, Christopher A. Elkins, and Romney M. Humphries. Framing bacterial genomics for public health (care). Journal of Clinical Microbiology, 59(12):e00135–21, 2021. doi: 10.1128/JCM.00135-21. URL https://journals.asm.org/doi/abs/10.1128/JCM.00135-21.

[25] Hilary D Marston, Dennis M Dixon, Jane M Knisely, Tara N Palmore, and Anthony S Fauci. Antimicrobial resistance. Jama, 316(11):1193–1204, 2016.

[26] Andrew G McArthur and Kara K Tsang. Antimicrobial resistance surveillance in the genomic age. Annals of the New York Academy of Sciences, 1388(1): 78–91, 2017.

[27] Chang-Hua Chen, Li-Chen Lin, Yu-Jun Chang, Yu-Min Chen, Chin-Yen Chang, and Chieh-Chen Huang. Infection control programs and antibiotic control programs to limit transmission of multi-drug resistant acinetobacter baumannii infections: evolution of old problems and new challenges for institutes. International journal of environmental research and public health, 12 (8):8871–8882, 2015.

[28] Francis S Codjoe and Eric S Donkor. Carbapenem resistance: a review. Medical Sciences, 6(1):1, 2017.

[29] Manar Ali Abushaheen, Amal Jamil Fatani, Mohammed Alosaimi, Wael Mansy, Merin George, Sadananda Acharya, Sanjay Rathod, Darshan Devang Divakar, Chitra Jhugroo, Sajith Vellappally, et al. Antimicrobial resistance, mechanisms and its clinical significance. Disease-a-Month, 66(6):100971, 2020.

[30] Fred C Tenover. Mechanisms of antimicrobial resistance in bacteria. The American journal of medicine, 119(6):S3–S10, 2006.

[31] Wanda C Reygaert. An overview of the antimicrobial resistance mechanisms of bacteria. AIMS microbiology, 4(3):482, 2018.

[32] Kelly L Wyres, Margaret Lam, and Kathryn E Holt. Population genomics of klebsiella pneumoniae. Nature Reviews Microbiology, 18(6):344–359, 2020.

[33] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3):443–453, 1970.

[34] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. Journal of Molecular Biology, 147(1):195–197, Mar 1981. ISSN 0022-2836. doi: 10.1016/0022-2836(81)90087-5. URL `http://dx.doi.org/10.1016/0022-2836(81)90087-5`.

[35] Dana Shapira and James A Storer. Edit distance with move operations. In Annual Symposium on Combinatorial Pattern Matching, pages 85–98. Springer, 2002.

[36] Arthur L Delcher, Simon Kasif, Robert D Fleischmann, Jeremy Peterson, Owen White, and Steven L Salzberg. Alignment of whole genomes. Nucleic acids research, 27(11):2369–2376, 1999.

[37] Lusheng Wang and Tao Jiang. On the complexity of multiple sequence alignment. Journal of computational biology, 1(4):337–348, 1994.

[38] Aaron CE Darling, Bob Mau, Frederick R Blattner, and Nicole T Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research, 14(7):1394–1403, 2004.

[39] Samuel V Angiuoli and Steven L Salzberg. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics, 27(3):334–342, 2011.

[40] Benedict Paten, Dent Earl, Ngan Nguyen, Mark Diekhans, Daniel Zerbino, and David Haussler. Cactus: Algorithms for genome multiple sequence alignment. Genome research, 21(9):1512–1528, 2011.

[41] Dent Earl, Ngan Nguyen, Glenn Hickey, Robert S Harris, Stephen Fitzgerald, Kathryn Beal, Igor Seledtsov, Vladimir Molodtsov, Brian J Raney, Hiram Clawson, et al. Alignathon: a competitive assessment of whole-genome alignment methods. Genome research, 24(12):2077–2089, 2014.

[42] Todd J Treangen, Brian D Ondov, Sergey Koren, and Adam M Phillippy. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome biology, 15(11):524, 2014.

[43] Steve Davis, James B Pettengill, Yan Luo, Justin Payne, Al Shpuntoff, Hugh Rand, and Errol Strain. Cfsan snp pipeline: an automated method for constructing snp matrices from next-generation sequence data. PeerJ Computer Science, 1:e20, 2015.

[44] Shea N Gardner and Barry G Hall. When whole-genome alignments just won't work: ksnp v2 software for alignment-free snp discovery and phylogenetics of hundreds of microbial genomes. PloS one, 8(12):e81760, 2013.

[45] Pascal Lapierre and J Peter Gogarten. Estimating the size of the bacterial pan-genome. Trends in genetics, 25(3):107–110, 2009.

[46] Hervé Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. Comparative genomics: the bacterial pan-genome. Current opinion in microbiology, 11(5): 472–477, 2008.

[47] Chad Laing, Cody Buchanan, Eduardo N Taboada, Yongxiang Zhang, Andrew Kropinski, Andre Villegas, James E Thomas, and Victor PJ Gannon. Pan-genome sequence analysis using panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC bioinformatics, 11(1):461, 2010.

[48] Yongbing Zhao, Jiayan Wu, Junhui Yang, Shixiang Sun, Jingfa Xiao, and Jun Yu. Pgap: pan-genomes analysis pipeline. Bioinformatics, 28(3):416–418, 2012.

[49] Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew TG Holden, Maria Fookes, Daniel Falush, Jacqueline A Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31(22):3691–3693, 2015.

[50] Neelam Goel, Shailendra Singh, and Trilok Chand Aseri. A review of soft computing techniques for gene prediction. International Scholarly Research Notices, 2013, 2013.

[51] Esther Zander, Agnieszka Chmielarczyk, Piotr Heczko, Harald Seifert, and Paul G Higgins. Conversion of oxa-66 into oxa-82 in clinical acinetobacter baumannii isolates and association with altered carbapenem susceptibility. Journal of Antimicrobial Chemotherapy, 68(2):308–311, 2013.

[52] Jennifer L Seffernick, Mervyn L de Souza, Michael J Sadowsky, and Lawrence P Wackett. Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. Journal of bacteriology, 183(8): 2405–2410, 2001.

[53] Harry A Thorpe, Sion C Bayliss, Samuel K Sheppard, and Edward J Feil. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. Gigascience, 7(4):giy015, 2018.

[54] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de bruijn graphs. Nature genetics, 44(2):226–232, 2012.

[55] Shoshana Marcus, Hayan Lee, and Michael C Schatz. Splitmem: a graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics, 30(24): 3476–3483, 2014.

[56] Ilia Minkin, Son Pham, and Paul Medvedev. Twopaco: An efficient algorithm to build the compacted de bruijn graph from many complete genomes. Bioinformatics, 33(24):4024–4032, 2017.

[57] Jamshed Khan and Rob Patro. Cuttlefish: fast, parallel and low-memory compaction of de bruijn graphs from large-scale genome collections. Bioinformatics, 37(Supplement_1):i177–i186, 2021.

[58] Brian P Alcock, Amogelang R Raphenya, Tammy TY Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, et al. Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic acids research, 48(D1):D517–D525, 2020.

[59] Bo Liu and Mihai Pop. Ardb—antibiotic resistance genes database. Nucleic acids research, 37(suppl_1):D443–D447, 2009.

[60] Sushim Kumar Gupta, Babu Roshan Padmanabhan, Seydina M Diene, Rafael Lopez-Rojas, Marie Kempf, Luce Landraud, and Jean-Marc Rolain. Argannot, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrobial agents and chemotherapy, 58(1):212–220, 2014.

[61] Gustavo Arango-Argoty, Emily Garner, Amy Pruden, Lenwood S Heath, Peter Vikesland, and Liqing Zhang. Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome, 6 (1):1–15, 2018.

[62] Alice R Wattam, James J Davis, Rida Assaf, Sébastien Boisvert, Thomas Brettin, Christopher Bun, Neal Conrad, Emily M Dietrich, Terry Disz, Joseph L Gabbard, et al. Improvements to patric, the all-bacterial bioinformatics database and analysis resource center. Nucleic acids research, 45(D1):D535–D542, 2017.

[63] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. The American Journal of Human Genetics, 101(1):5–22, 2017.

[64] Baiqiang Liang, Hongrong Ding, Lianfang Huang, Haiqing Luo, and Xiao Zhu. Gwas in cancer: progress and challenges. Molecular Genetics and Genomics, 295(3):537–561, 2020.

[65] John A Lees, Minna Vehkala, Niko Välimäki, Simon R Harris, Claire Chewapreecha, Nicholas J Croucher, Pekka Marttinen, Mark R Davies, Andrew C Steer, Steven YC Tong, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nature communications, 7(1):1–8, 2016.

[66] Magali Jaillard, Leandro Lima, Maud Tournoud, Pierre Mahé, Alex Van Belkum, Vincent Lacroix, and Laurent Jacob. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. PLoS genetics, 14(11):e1007758, 2018.

[67] N. Ricker, H. Qian, and R.R. Fulthorpe. The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. Genomics, 100 (3):167–175, 2012. ISSN 0888-7543. doi: https://doi.org/10.1016/j.ygeno.2012.06.009. URL https://www.sciencedirect.com/science/article/pii/S0888754312001279.

[68] Claire Bertelli, Keith E Tilley, and Fiona SL Brinkman. Microbial genomic island discovery, visualization and analysis. Briefings in bioinformatics, 20(5): 1685–1698, 2019.

[69] Mario Juhas, Jan Roelof Van Der Meer, Muriel Gaillard, Rosalind M Harding, Derek W Hood, and Derrick W Crook. Genomic islands: tools of bacterial

horizontal gene transfer and evolution. FEMS microbiology reviews, 33(2): 376–393, 2009.

[70] Mark D Adams, Brian Bishop, and Meredith S Wright. Quantitative assessment of insertion sequence impact on bacterial genome architecture. Microbial genomics, 2(7), 2016.

[71] Benjamin A Evans, Ahmed Hamouda, and Sebastian GB Amyes. The rise of carbapenem-resistant acinetobacter baumannii. Current pharmaceutical design, 19(2):223–238, 2013.

[72] Yonatan H Grad and Marc Lipsitch. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. Genome biology, 15(11):1–14, 2014.

[73] Nina Van Goethem, Tine Descamps, Brecht Devleesschauwer, Nancy HC Roosens, Nele AM Boon, Herman Van Oyen, and Annie Robert. Status and potential of bacterial genomics for public health practice: a scoping review. Implementation Science, 14(1):1–16, 2019.

[74] Kalliopi Rantsiou, Sophia Kathariou, Annet Winkler, Panos Skandamis, Manuel Jimmy Saint-Cyr, Katia Rouzeau-Szynalski, and Alejandro Amézquita. Next generation microbiological risk assessment: opportunities of whole genome sequencing (wgs) for foodborne pathogen surveillance, source tracking and risk assessment. International journal of food microbiology, 287: 3–9, 2018.

[75] Ji-Yeon Hyeon, Shaoting Li, David A. Mann, Shaokang Zhang, Zhen Li, Yi Chen, Xiangyu Deng, and Charles M. Dozois. Quasimetagenomics-based and real-time-sequencing-aided detection and subtyping of salmonella enterica from food samples. Applied and Environmental Microbiology, 84(4):e02340–17, 2018. doi: 10.1128/AEM.02340-17. URL https://journals.asm.org/doi/abs/10.1128/AEM.02340-17.

[76] Amina Baraketi, Stephane Salmieri, and Monique Lacroix. Foodborne pathogens detection: persevering worldwide challenge. Biosensing technologies for the detection of pathogens—a prospective way for rapid analysis. IntechOpen, Rijeka, Croatia, pages 53–72, 2018.

[77] Surendra Rasamsetti, Mark Berrang, Nelson A. Cox, and Nikki W. Shariat. Selective pre-enrichment method to lessen time needed to recover salmonella from commercial poultry processing samples. Food Microbiology, 99: 103818, 2021. ISSN 0740-0020. doi: https://doi.org/10.1016/j.fm.2021.103818. URL https://www.sciencedirect.com/science/article/pii/S0740002021000836.

[78] Andrea R Ottesen, Antonio Gonzalez, Rebecca Bell, Caroline Arce, Steven Rideout, Marc Allard, Peter Evans, Errol Strain, Steven Musser, Rob Knight,

et al. Co-enriching microflora associated with culture based methods to detect salmonella from tomato phyllosphere. PloS one, 8(9):e73079, 2013.

[79] Eli L Moss, Dylan G Maghini, and Ami S Bhatt. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nature biotechnology, 38(6):701–707, 2020.

[80] Andrea Ottesen, Padmini Ramachandran, Yi Chen, Eric Brown, Elizabeth Reed, and Errol Strain. Quasimetagenomic source tracking of listeria monocytogenes from naturally contaminated ice cream. BMC Infectious Diseases, 20(1):1–6, 2020.

[81] Jo Handelsman. Metagenomics: application of genomics to uncultured microorganisms. Microbiology and molecular biology reviews, 68(4):669–685, 2004.

[82] Asiya Nazir. Review on metagenomics and its applications. Imp J Intersd Res, 2(10), 2016.

[83] Patrick Lorenz and Jürgen Eck. Metagenomics and industrial applications. Nature Reviews Microbiology, 3(6):510–516, 2005.

[84] Aravind Madhavan, Raveendran Sindhu, Parameswaran Binod, Rajeev K Sukumaran, and Ashok Pandey. Strategies for design of improved biocatalysts for industrial applications. Bioresource technology, 245:1304–1313, 2017.

[85] Mihai Pop, Alan W Walker, Joseph Paulson, Brianna Lindsay, Martin Antonio, M Anowar Hossain, Joseph Oundo, Boubou Tamboura, Volker Mai, Irina Astrovskaya, et al. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. Genome biology, 15(6):1–12, 2014.

[86] Jacquelyn S Meisel, Daniel J Nasko, Brian Brubach, Victoria Cepeda-Espinoza, Jessica Chopyk, Héctor Corrada-Bravo, Marcus Fedarko, Jay Ghurye, Kiran Javkar, Nathan D Olson, et al. Current progress and future opportunities in applications of bioinformatics for biodefense and pathogen detection: report from the winter mid-atlantic microbiome meet-up, college park, md, january 10, 2018, 2018.

[87] Vincenzo Pennone, José F Cobo-Díaz, Miguel Prieto, and Avelino Alvarez-Ordóñez. Application of genomics and metagenomics to improve food safety based on an enhanced characterisation of antimicrobial resistance. Current Opinion in Food Science, 43:183–188, 2022. ISSN 2214-7993. doi: https://doi.org/10.1016/j.cofs.2021.12.002. URL https://www.sciencedirect.com/science/article/pii/S2214799321001624.

[88] Florian Plaza Onate, Jean-Michel Batto, Catherine Juste, Jehane Fadlallah, Cyrielle Fougeroux, Doriane Gouas, Nicolas Pons, Sean Kennedy, Florence

Levenez, Joel Dore, et al. Quality control of microbiota metagenomics by k-mer analysis. BMC genomics, 16(1):1–10, 2015.

[89] J. Paul Brooks. Challenges for case-control studies with microbiome data. Annals of Epidemiology, 26(5):336–341.e1, 2016. ISSN 1047-2797. doi: https://doi.org/10.1016/j.annepidem.2016.03.009. URL https://www.sciencedirect.com/science/article/pii/S1047279716300795. The Microbiome and Epidemiology.

[90] Tanya M Monaghan, Tim J Sloan, Stephen R Stockdale, Adam M Blanchard, Richard D Emes, Mark Wilcox, Rima Biswas, Rupam Nashine, Sonali Manke, Jinal Gandhi, et al. Metagenomics reveals impact of geography and acute diarrheal disease on the central indian human gut microbiome. Gut Microbes, pages 1–24, 2020.

[91] Tommi Vatanen, Damian R Plichta, Juhi Somani, Philipp C Münch, Timothy D Arthur, Andrew Brantley Hall, Sabine Rudolf, Edward J Oakeley, Xiaobo Ke, Rachel A Young, et al. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nature microbiology, 4(3):470–479, 2019.

[92] Romney M Humphries and Andrea J Linscott. Laboratory diagnosis of bacterial gastroenteritis. Clinical microbiology reviews, 28(1):3–31, 2015.

[93] Philip P Ahern, Jeremiah J Faith, and Jeffrey I Gordon. Mining the human gut microbiota for effector strains that shape the immune system. Immunity, 40(6):815–823, 2014.

[94] Stephen Nayfach and Katherine S. Pollard. Toward accurate and quantitative comparative metagenomics. Cell, 166(5):1103–1116, 2016. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2016.08.007. URL https://www.sciencedirect.com/science/article/pii/S009286741631056X.

[95] Pajau Vangay, Abigail J Johnson, Tonya L Ward, Gabriel A Al-Ghalith, Robin R Shields-Cutler, Benjamin M Hillmann, Sarah K Lucas, Lalit K Beura, Emily A Thompson, Lisa M Till, et al. Us immigration westernizes the human gut microbiome. Cell, 175(4):962–972, 2018.

[96] S Commichaux, K Javkar, HS Muralidharan, P Ramachandran, A Ottesen, H Rand, and M Pop. Taxatarget: Fast, sensitive, and precise classification of microeukaryotes in metagenomic data. 2021. doi: 10.21203/rs.3.rs-1186624/v3.

[97] Gherman Uritskiy and Jocelyne DiRuggiero. Applying genome-resolved metagenomics to deconvolute the halophilic microbiome. Genes, 10(3):220, 2019.

[98] Florian P Breitwieser, Jennifer Lu, and Steven L Salzberg. A review of methods and databases for metagenomic classification and assembly. Briefings in bioinformatics, 20(4):1125–1136, 2019.

[99] Seth Commichaux, Kiran Javkar, Padmini Ramachandran, Niranjan Nagarajan, Denis Bertrand, Yi Chen, Elizabeth Reed, Narjol Gonzalez-Escalona, Errol Strain, Hugh Rand, et al. Evaluating the accuracy of listeria monocytogenes assemblies from quasimetagenomic samples using long and short reads. BMC genomics, 22(1):1–18, 2021. doi: 10.1186/s12864-021-07702-2.

[100] Jim O'Neill. Tackling drug-resistant infections globally: final report and recommendations. 2016.

[101] Mohammad Hamidian and Steven J Nigro. Emergence, molecular mechanisms and global spread of carbapenem-resistant acinetobacter baumannii. Microbial genomics, 5(10), 2019.

[102] Helen W Boucher, George H Talbot, John S Bradley, John E Edwards, David Gilbert, Louis B Rice, Michael Scheld, Brad Spellberg, and John Bartlett. Bad bugs, no drugs: no eskape! an update from the infectious diseases society of america. Clinical infectious diseases, 48(1):1–12, 2009.

[103] Evelina Tacconelli, Elena Carrara, Alessia Savoldi, Stephan Harbarth, Marc Mendelson, Dominique L Monnet, Céline Pulcini, Gunnar Kahlmeter, Jan Kluytmans, Yehuda Carmeli, et al. Discovery, research, and development of new antibiotics: the who priority list of antibiotic-resistant bacteria and tuberculosis. The Lancet Infectious Diseases, 18(3):318–327, 2018.

[104] Ignasi Roca Subirà, Paula Espinal, Xavier Vila-Farrés, and Jordi Vila Estapé. The acinetobacter baumannii oxymoron: commensal hospital dweller turned pan-drug-resistant menace. Frontiers in microbiology, 3:148, 2012.

[105] Jennifer Nowak, E Zander, Danuta Stefanik, Paul G Higgins, Ignasi Roca, Jordi Vila, Michael J McConnell, José Miguel Cisneros, Harald Seifert, and MagicBullet Working Group WP4. High incidence of pandrug-resistant acinetobacter baumannii isolates collected from patients with ventilator-associated pneumonia in greece, italy and spain as part of the magicbullet clinical trial. Journal of Antimicrobial Chemotherapy, 72(12):3277–3282, 2017.

[106] Anton Y Peleg, Harald Seifert, and David L Paterson. Acinetobacter baumannii: emergence of a successful pathogen. Clinical microbiology reviews, 21(3): 538–582, 2008.

[107] Raffaele Zarrilli, Spyros Pournaras, Maria Giannouli, and Athanassios Tsakris. Global evolution of multidrug-resistant acinetobacter baumannii clonal lineages. International journal of antimicrobial agents, 41(1):11–19, 2013.

[108] Eli Ben-Chetrit, Yonit Wiener-Well, Emil Lesho, Puah Kopuit, Chaya Broyer, Liora Bier, Marc V Assous, Shmuel Benenson, Matan J Cohen, Patrick T McGann, et al. An intervention to control an icu outbreak of carbapenem-resistant acinetobacter baumannii: long-term impact for the icu and hospital. Critical Care, 22(1):1–10, 2018.

[109] Laurent Poirel and P Nordmann. Carbapenem resistance in acinetobacter baumannii: mechanisms and epidemiology. Clinical Microbiology and Infection, 12(9):826–836, 2006.

[110] Mariana Pagano, Andreza Francisco Martins, and Afonso Luis Barth. Mobile genetic elements related to carbapenem resistance in acinetobacter baumannii. brazilian journal of microbiology, 47:785–792, 2016.

[111] Benjamin A Evans and Sebastian GB Amyes. Oxa β-lactamases. Clinical microbiology reviews, 27(2):241–263, 2014.

[112] Carsten Kröger, Stefani C Kary, Kristina Schauer, and Andrew DS Cameron. Genetic regulation of virulence and antibiotic resistance in acinetobacter baumannii. Genes, 8(1):12, 2016.

[113] Stephane Corvec, Nathalie Caroff, Eric Espaze, Cecile Giraudeau, Henri Drugeon, and Alain Reynaud. Ampc cephalosporinase hyperproduction in acinetobacter baumannii clinical strains. Journal of antimicrobial chemotherapy, 52(4):629–635, 2003.

[114] Jane F Turton, M Elaina Ward, Neil Woodford, Mary E Kaufmann, Rachel Pike, David M Livermore, and Tyrone L Pitt. The role of is aba1 in expression of oxa carbapenemase genes in acinetobacter baumannii. FEMS microbiology letters, 258(1):72–77, 2006.

[115] Emma C Schroder, Zachary L Klamer, Aysegul Saral, Kyle A Sugg, Cynthia M June, Troy Wymore, Agnieszka Szarecka, and David A Leonard. Clinical variants of the native class d β-lactamase of acinetobacter baumannii pose an emerging threat through increased hydrolytic activity against carbapenems. Antimicrobial agents and chemotherapy, 60(10):6155–6164, 2016.

[116] Nuno T Antunes, Toni L Lamoureaux, Marta Toth, Nichole K Stewart, Hilary Frase, and Sergei B Vakulenko. Class d β-lactamases: are they all carbapenemases? Antimicrobial agents and chemotherapy, 58(4):2119–2125, 2014.

[117] Claire Héritier, Laurent Poirel, Pierre-Edouard Fournier, Jean-Michel Claverie, Didier Raoult, and Patrice Nordmann. Characterization of the naturally occurring oxacillinase of acinetobacter baumannii. Antimicrobial agents and chemotherapy, 49(10):4174–4179, 2005.

[118] Yuiko Takebayashi, Jacqueline Findlay, Kate J Heesom, Philip J Warburton, Matthew B Avison, and Benjamin A Evans. Variability in carbapenemase

activity of intrinsic oxaab (oxa-51-like) β-lactamase enzymes in acinetobacter baumannii. Journal of Antimicrobial Chemotherapy, 76(3):587–595, 2021.

[119] Claire Héritier, Laurent Poirel, Thierry Lambert, and Patrice Nordmann. Contribution of acquired carbapenem-hydrolyzing oxacillinases to carbapenem resistance in acinetobacter baumannii. Antimicrobial agents and chemotherapy, 49(8):3198–3202, 2005.

[120] Karyne Rangel Carvalho, Ana Paula D'Alincourt Carvalho-Assef, Lia Galvão dos Santos, Maria José Félix Pereira, and Marise Dutra Asensi. Occurrence of blaoxa-23 gene in imipenem-susceptible acinetobacter baumannii. Memórias do Instituto Oswaldo Cruz, 106:505–506, 2011.

[121] TW Boo and B Crowley. Detection of blaoxa-58 and blaoxa-23-like genes in carbapenem-susceptible acinetobacter clinical isolates: should we be concerned? Journal of Medical Microbiology, 58(6):839–841, 2009.

[122] Lalena Wallace, Sean C Daugherty, Sushma Nagaraj, J Kristie Johnson, Anthony D Harris, and David A Rasko. Use of comparative genomics to characterize the diversity of acinetobacter baumannii surveillance isolates in a health care institution. Antimicrobial agents and chemotherapy, 60(10):5933–5941, 2016.

[123] Rodrigo Cayô, María-Cruz Rodríguez, Paula Espinal, Felipe Fernández-Cuenca, Alain A Ocampo-Sosa, Alvaro Pascual, Juan A Ayala, Jordi Vila, and Luis Martínez-Martínez. Analysis of genes encoding penicillin-binding proteins in clinical isolates of acinetobacter baumannii. Antimicrobial agents and chemotherapy, 55(12):5907–5913, 2011.

[124] Khadidja Khorsi, Yamina Messai, Moufida Hamidi, Houria Ammari, and Rabah Bakour. High prevalence of multidrug-resistance in acinetobacter baumannii and dissemination of carbapenemase-encoding genes blaoxa-23-like, blaoxa-24-like and blandm-1 in algiers hospitals. Asian Pacific journal of tropical medicine, 8(6):438–446, 2015.

[125] María A Mussi, Adriana S Limansky, and Alejandro M Viale. Acquisition of resistance to carbapenems in multidrug-resistant clinical strains of acinetobacter baumannii: natural insertional inactivation of a gene encoding a member of a novel family of β-barrel outer membrane proteins. Antimicrobial agents and chemotherapy, 49(4):1432–1440, 2005.

[126] Hye Won Jeong, Hee Jin Cheong, Woo Joo Kim, Min Ja Kim, Ki-Joon Song, Jin-Won Song, H Stanley Kim, and Kyoung Ho Roh. Loss of the 29-kilodalton outer membrane protein in the presence of oxa-51–like enzymes in acinetobacter baumannii is associated with decreased imipenem susceptibility. Microbial Drug Resistance, 15(3):151–158, 2009.

[127] Su-ying Zhao, Dong-yang Jiang, Peng-cheng Xu, Yi-kai Zhang, Heng-fang Shi, Hui-ling Cao, and Qian Wu. An investigation of drug-resistant acinetobacter baumannii infections in a comprehensive hospital of east china. Annals of clinical microbiology and antimicrobials, 14(1):1–8, 2015.

[128] Jitendra Vashist, Vishvanath Tiwari, Rituparna Das, Arti Kapil, and Moganty R Rajeswari. Analysis of penicillin-binding proteins (pbps) in carbapenem resistant acinetobacter baumannii. The Indian journal of medical research, 133(3):332, 2011.

[129] Jane Hawkey, David B Ascher, Louise M Judd, Ryan R Wick, Xenia Kostoulias, Heather Cleland, Denis W Spelman, Alex Padiglione, Anton Y Peleg, and Kathryn E Holt. Evolution of carbapenem resistance in acinetobacter baumannii during a prolonged infection. Microbial genomics, 4(3), 2018.

[130] Wei Jia, Caiyun Li, Haiyun Zhang, Gang Li, Xiaoming Liu, and Jun Wei. Prevalence of genes of oxa-23 carbapenemase and adeabc efflux pump associated with multidrug resistance of acinetobacter baumannii isolates in the icu of a comprehensive hospital of northwestern china. International Journal of Environmental Research and Public Health, 12(8):10079–10092, 2015.

[131] Kai-Chih Chang, Han-Yueh Kuo, Chuan Yi Tang, Cheng-Wei Chang, Chia-Wei Lu, Chih-Chin Liu, Huei-Ru Lin, Kuan-Hsueh Chen, and Ming-Li Liou. Transcriptome profiling in imipenem-selected acinetobacter baumannii. BMC genomics, 15(1):1–13, 2014.

[132] Fei Liu, Yuying Zhu, Yong Yi, Na Lu, Baoli Zhu, and Yongfei Hu. Comparative genomic analysis of acinetobacter baumannii clinical isolates reveals extensive genomic variation and diverse antibiotic resistance determinants. BMC genomics, 15(1):1–14, 2014.

[133] Chris Rowe Taitt, Tomasz A Leski, Michael G Stockelman, David W Craft, Daniel V Zurawski, Benjamin C Kirkup, and Gary J Vora. Antimicrobial resistance determinants in acinetobacter baumannii isolates taken from military treatment facilities. Antimicrobial agents and chemotherapy, 58(2):767–781, 2014.

[134] Nicola C Gordon and David W Wareham. Multidrug-resistant acinetobacter baumannii: mechanisms of virulence and resistance. International journal of antimicrobial agents, 35(3):219–226, 2010.

[135] Simona Bratu, David Landman, Don Antonio Martin, Claudiu Georgescu, and John Quale. Correlation of antimicrobial resistance with β-lactamases, the ompa-like porin, and efflux pumps in clinical isolates of acinetobacter baumannii endemic to new york city. Antimicrobial agents and chemotherapy, 52(9):2999–3005, 2008.

[136] Dexi Bi, Ruting Xie, Jiayi Zheng, Huiqiong Yang, Xingchen Zhu, Hong-Yu Ou, and Qing Wei. Large-scale identification of abar-type genomic islands in acinetobacter baumannii reveals diverse insertion sites and clonal lineage-specific antimicrobial resistance gene profiles. Antimicrobial agents and chemotherapy, 63(4):e02526–18, 2019.

[137] Laure Diancourt, Virginie Passet, Alexandr Nemec, Lenie Dijkshoorn, and Sylvain Brisse. The population structure of acinetobacter baumannii: expanding multiresistant clones from an ancestral susceptible genetic pool. PloS one, 5(4):e10034, 2010.

[138] Philip M Ashton, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'grady. Minion nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nature biotechnology, 33(3):296–300, 2015.

[139] Andrea M Hujer, Paul G Higgins, Susan D Rudin, Genevieve L Buser, Steven H Marshall, Kyriaki Xanthopoulou, Harald Seifert, Laura J Rojas, T Nicholas Domitrovic, P Maureen Cassidy, et al. Nosocomial outbreak of extensively drug-resistant acinetobacter baumannii isolates containing bla oxa-237 carried on a plasmid. Antimicrobial agents and chemotherapy, 61(11): e00797–17, 2017.

[140] Henan Li, Fei Liu, Yawei Zhang, Xiaojuan Wang, Chunjiang Zhao, Hongbin Chen, Feifei Zhang, Baoli Zhu, Yongfei Hu, and Hui Wang. Evolution of carbapenem-resistant acinetobacter baumannii revealed through whole-genome sequencing and comparative genomic analysis. Antimicrobial agents and chemotherapy, 59(2):1168–1176, 2015.

[141] Saranya Vijayakumar, Chand Wattal, JK Oberoi, Sanjay Bhattacharya, Karthick Vasudevan, Shalini Anandan, Kamini Walia, and Balaji Veeraraghavan. Insights into the complete genomes of carbapenem-resistant acinetobacter baumannii harbouring bla oxa-23, bla oxa-420 and bla ndm-1 genes using a hybrid-assembly approach. Access microbiology, 2(8), 2020.

[142] David L Lin, German M Traglia, Rachel Baker, David J Sherratt, Maria Soledad Ramirez, and Marcelo E Tolmasky. Functional analysis of the acinetobacter baumannii xerc and xerd site-specific recombinases: potential role in dissemination of resistance genes. Antibiotics, 9(7):405, 2020.

[143] Arnaud Gutierrez, Marina Elez, Olivier Clermont, Erick Denamur, and Ivan Matic. Escherichia coli yafp protein modulates dna damaging property of the nitroaromatic compounds. Nucleic acids research, 39(10):4192–4201, 2011.

[144] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. Spades: a new genome assembly algorithm and its

applications to single-cell sequencing. Journal of computational biology, 19 (5):455–477, 2012.

[145] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. Quast: quality assessment tool for genome assemblies. Bioinformatics, 29 (8):1072–1075, 2013.

[146] Torsten Seemann. mlst. `https://github.com/tseemann/mlst`.

[147] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14):2068–2069, 2014.

[148] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22 (13):1658–1659, 2006.

[149] David L Lin, German M Traglia, Rachel Baker, David J Sherratt, Maria Soledad Ramirez, and Marcelo E Tolmasky. Functional analysis of the acinetobacter baumannii xerc and xerd site-specific recombinases: potential role in dissemination of resistance genes. Antibiotics, 9(7):405, 2020.

[150] Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, et al. Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. Nature methods, 10(6):563–569, 2013.

[151] Martin Hunt, Nishadi De Silva, Thomas D Otto, Julian Parkhill, Jacqueline A Keane, and Simon R Harris. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome biology, 16(1):1–10, 2015.

[152] Jay Ghurye, Todd Treangen, Marcus Fedarko, W Judson Hervey, and Mihai Pop. Metacarvel: linking assembly graph motifs to biological variants. Genome biology, 20(1):174, 2019.

[153] Derrick Joel Zwickl. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. the University of Texas at Austin, 2006.

[154] Kiran Javkar, Hugh Rand, Errol Strain, and Mihai Pop. Prawns: Pan-genome representation of a large number of whole genomes. `https://github.com/KiranJavkar/PRAWNS`, 2022.

[155] Valentin Zulkower and Susan Rosser. Dna features viewer, a sequence annotation formatting and plotting library for python. Bioinformatics, 2020.

[156] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. Journal of molecular biology, 215 (3):403–410, 1990.

[157] Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L Madden. Ncbi blast: a better web interface. Nucleic acids research, 36(suppl_2):W5–W9, 2008.

[158] Guillaume Marçais, Arthur L Delcher, Adam M Phillippy, Rachel Coston, Steven L Salzberg, and Aleksey Zimin. Mummer4: A fast and versatile genome alignment system. PLoS computational biology, 14(1):e1005944, 2018.

[159] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods, 17(3):261–272, 2020.

[160] SK Dritz, AF Back, et al. Shigella enteritis venereally transmitted. New England Journal of Medicine, 291(22), 1974.

[161] Max Bader, AHB Pedersen, ROBERT WILLIAMS, JEAN SPEARMAN, and HERB ANDERSON. Venereal transmission of shigellosis in seattle-king county. Sexually transmitted diseases, pages 89–91, 1977.

[162] LEWIS M Drusin, GAIL Genvert, BARBARA Topf-Olstein, and ELLEN Levy-Zombek. Shigellosis. another sexually transmitted disease? Sexually Transmitted Infections, 52(5):348–350, 1976.

[163] Tomás J Aragón, Duc J Vugia, Sue Shallow, Michael C Samuel, Arthur Reingold, Frederick J Angulo, and Williamson Z Bradford. Case-control study of shigellosis in san francisco: the role of sexual transmission and hiv infection. Clinical Infectious Diseases, 44(3):327–334, 2007.

[164] Anna Bowen, Dana Eikmeier, Pamela Talley, Alicia Siston, Shamika Smith, Jacqueline Hurd, Kirk Smith, Fe Leano, Amelia Bicknese, J Corbin Norton, et al. Outbreaks of shigella sonnei infection with decreased susceptibility to azithromycin among men who have sex with men—chicago and metropolitan minneapolis-st. paul, 2014. MMWR. Morbidity and mortality weekly report, 64(21):597, 2015.

[165] Jonas Z Hines. Notes from the field: shigellosis outbreak among men who have sex with men and homeless persons—oregon, 2015–2016. MMWR. Morbidity and Mortality Weekly Report, 65, 2016.

[166] Anna Bowen, Julian Grass, Amelia Bicknese, Davina Campbell, Jacqueline Hurd, and Robert D Kirkcaldy. Elevated risk for antimicrobial drug–resistant shigella infection among men who have sex with men, united states, 2011–2015. Emerging infectious diseases, 22(9):1613, 2016.

[167] Richard N Danila, Dana L Eikmeier, Trisha J Robinson, Allison La Pointe, and Aaron S DeVries. Two concurrent enteric disease outbreaks among men who have sex with men, minneapolis–st paul area. Clinical Infectious Diseases, 59(7):987–989, 2014.

[168] Hsiao-Han Chang, Ted Cohen, Yonatan H Grad, William P Hanage, Thomas F O'Brien, and Marc Lipsitch. Origin and proliferation of multiple-drug resistance in bacterial pathogens. Microbiology and Molecular Biology Reviews, 79(1):101–116, 2015.

[169] A-P Magiorakos, A Srinivasan, Roberta B Carey, Yehuda Carmeli, ME Falagas, CG Giske, S Harbarth, JF Hindler, Gunnar Kahlmeter, Barbro Olsson-Liljequist, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. Clinical microbiology and infection, 18(3): 268–281, 2012.

[170] Kenya Murray, Vasudha Reddy, John S Kornblum, HaeNa Waechter, Ludwin F Chicaiza, Inessa Rubinstein, Sharon Balter, Sharon K Greene, Sarah L Braunstein, Jennifer L Rakeman, et al. Increasing antibiotic resistance in shigella spp. from infected new york city residents, new york, usa. Emerging infectious diseases, 23(2):332, 2017.

[171] Katherine E Heiman, Maria Karlsson, Julian Grass, Becca Howie, Robert D Kirkcaldy, Barbara Mahon, John T Brooks, and Anna Bowen. Shigella with decreased susceptibility to azithromycin among men who have sex with men—united states, 2002–2013. MMWR. Morbidity and Mortality Weekly Report, 63(6):132, 2014.

[172] Karl C Klontz and Nalini Singh. Treatment of drug-resistant shigella infections. Expert review of anti-infective therapy, 13(1):69–80, 2015.

[173] Jean B Patel, FR Cockerill, and Patricia A Bradford. Performance standards for antimicrobial susceptibility testing: twenty-fifth informational supplement. 2015.

[174] Varvara K Kozyreva, Guillaume Jospin, Alexander L Greninger, James P Watt, Jonathan A Eisen, and Vishnu Chaturvedi. Recent outbreaks of shigellosis in california caused by two distinct populations of shigella sonnei with either increased virulence or fluoroquinolone resistance. Msphere, 1(6):e00344–16, 2016.

[175] Kate S Baker, Timothy J Dallman, Nigel Field, Tristan Childs, Holly Mitchell, Martin Day, François-Xavier Weill, Sophie Lefèvre, Mathieu Tourdjman, Gwenda Hughes, et al. Horizontal antimicrobial resistance transfer drives epidemics of multiple shigella species. Nature communications, 9(1):1–10, 2018.

[176] Kate S Baker, Timothy J Dallman, Philip M Ashton, Martin Day, Gwenda Hughes, Paul D Crook, Victoria L Gilbart, Sandra Zittermann, Vanessa G Allen, Benjamin P Howden, et al. Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: a cross-sectional study. The Lancet infectious diseases, 15(8):913–921, 2015.

[177] Danielle J Ingle, Marion Easton, Mary Valcanis, Torsten Seemann, Jason C Kwong, Nicola Stephens, Glen P Carter, Anders Gonçalves da Silva, James Adamopoulos, Sarah L Baines, et al. Co-circulation of multidrug-resistant shigella among men who have sex with men in australia. Clinical Infectious Diseases, 69(9):1535–1544, 2019.

[178] Khadidja Yousfi, Christiane Gaudreau, Pierre A Pilon, Brigitte Lefebvre, Matthew Walker, Éric Fournier, Florence Doualla Bell, Christine Martineau, Jean Longtin, and Sadjia Bekal. Genetic mechanisms behind the spread of reduced susceptibility to azithromycin in shigella strains isolated from men who have sex with men in québec, canada. Antimicrobial Agents and Chemotherapy, 63(2):e01679–18, 2019.

[179] Marc W Allard, Errol Strain, David Melka, Kelly Bunning, Steven M Musser, Eric W Brown, and Ruth Timme. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. Journal of clinical microbiology, 54(8):1975–1983, 2016.

[180] Bala Swaminathan, Timothy J Barrett, Susan B Hunter, Robert V Tauxe, and CDC PulseNet Task Force. Pulsenet: the molecular subtyping network for foodborne bacterial disease surveillance, united states. Emerging infectious diseases, 7(3):382, 2001.

[181] Kate S Baker, Timothy J Dallman, Nigel Field, Tristan Childs, Holly Mitchell, Martin Day, François-Xavier Weill, Sophie Lefèvre, Mathieu Tourdjman, Gwenda Hughes, et al. Genomic epidemiology of shigella in the united kingdom shows transmission of pathogen sublineages and determinants of antimicrobial resistance. Scientific reports, 8(1):1–8, 2018.

[182] Fan Yang, Jian Yang, Xiaobing Zhang, Lihong Chen, Yan Jiang, Yongliang Yan, Xudong Tang, Jing Wang, Zhaohui Xiong, Jie Dong, et al. Genome dynamics and diversity of shigella species, the etiologic agents of bacillary dysentery. Nucleic acids research, 33(19):6445–6458, 2005.

[183] Cristóbal Almendros, Francisco JM Mojica, César Díez-Villaseñor, Noemí M Guzmán, and Jesús García-Martínez. Crispr-cas functional module exchange in escherichia coli. MBio, 5(1):e00767–13, 2014.

[184] Jaroslaw Zdziarski, Elzbieta Brzuszkiewicz, Björn Wullt, Heiko Liesegang, Dvora Biran, Birgit Voigt, Jenny Grönberg-Hernandez, Bryndis Ragnarsdottir, Michael Hecker, Eliora Z Ron, et al. Host imprints on bacterial genomes—rapid, divergent evolution in individual patients. PLoS pathogens, 6(8): e1001078, 2010.

[185] Sylvie Miquel, Eric Peyretaillade, Laurent Claret, Amélie De Vallée, Carole Dossat, Benoit Vacherie, El Hajji Zineb, Beatrice Segurens, Valerie Barbe, Pierre Sauvanet, et al. Complete genome sequence of crohn's disease-associated adherent-invasive e. coli strain lf82. PloS one, 5(9):e12714, 2010.

[186] John F Beltrami, R Luke Shouse, and Paul A Blake. Trends in infectious diseases and the male to female ratio: possible clues to changes in behavior among men who have sex with men. AIDS Education & Prevention, 17 (Supplement B):49–59, 2005.

[187] Piers Mook, D Gardiner, S Kanagarajah, M Kerac, Gwenda Hughes, Nigel Field, Noel McCarthy, C Rawlings, Ian Simms, C Lane, et al. Use of gender distribution in routine surveillance data to detect potential transmission of gastrointestinal infections among men who have sex with men in england. Epidemiology & Infection, 146(11):1468–1477, 2018.

[188] Stephen Baker and Hao Chung The. Recent insights into shigella: a major contributor to the global diarrhoeal disease burden. Current opinion in infectious diseases, 31(5):449, 2018.

[189] Mahbubur Rahman, Shereen Shoma, Harunur Rashid, AK Siddique, GB Nair, and DA Sack. Extended-spectrum β-lactamase-mediated third-generation cephalosporin resistance in shigella isolates in bangladesh. Journal of Antimicrobial Chemotherapy, 54(4):846–847, 2004.

[190] Neelam Taneja, Abhishek Mewara, Ajay Kumar, Garima Verma, and Meera Sharma. Cephalosporin-resistant shigella flexneri over 9 years (2001–09) in india. Journal of antimicrobial chemotherapy, 67(6):1347–1353, 2012.

[191] Wenli Zhang, Yanping Luo, Jingyun Li, Lan Lin, Yue Ma, Changqin Hu, Shaohong Jin, Lu Ran, and Shenghui Cui. Wide dissemination of multidrug-resistant shigella isolates in china. Journal of Antimicrobial Chemotherapy, 66(11):2527–2535, 2011.

[192] I-Fei Huang, Cheng-Hsun Chiu, Mei-Hui Wang, Chan-Yao Wu, Kai-Sheng Hsieh, and Christine C Chiou. Outbreak of dysentery associated with ceftriaxone-resistant shigella sonnei: first report of plasmid-mediated cmy-2-type ampc β-lactamase resistance in s. sonnei. Journal of Clinical Microbiology, 43(6):2608–2612, 2005.

[193] Nicole D Pecora, Ning Li, Marc Allard, Cong Li, Esperanza Albano, Mary Delaney, Andrea Dubois, Andrew B Onderdonk, and Lynn Bry. Genomically informed surveillance for carbapenem-resistant enterobacteriaceae in a health care system. MBio, 6(4):e01030–15, 2015.

[194] Kathleen Nudel, Xiaomin Zhao, Sankha Basu, Xiaoxi Dong, Maria Hoffmann, Michael Feldgarden, Marc Allard, Michael Klompas, and Lynn Bry. Genomics of corynebacterium striatum, an emerging multidrug-resistant pathogen of immunocompromised patients. Clinical Microbiology and Infection, 24(9): 1016–e7, 2018.

[195] Giuseppe Bertani. Studies on lysogenesis i: the mode of phage liberation by lysogenic escherichia coli. Journal of bacteriology, 62(3):293–300, 1951.

[196] Kyoung-Hee Choi, Ayush Kumar, and Herbert P Schweizer. A 10-min method for preparation of highly electrocompetent pseudomonas aeruginosa cells: application for dna fragment transfer between chromosomes and plasmid transformation. Journal of microbiological methods, 64(3):391–397, 2006.

[197] Mary Jane Ferraro. Performance standards for antimicrobial disk susceptibility tests. NCCLS, 2000.

[198] Matthew A Wikler. Methods for dilution antimicrobial susceptibility tests for bacteria that grow aerobically: approved standard. CLSI (NCCLS), 26: M7–A7, 2006.

[199] Nicole Pecora, Xiaomin Zhao, Kathleen Nudel, Maria Hoffmann, Ning Li, Andrew B Onderdonk, Deborah Yokoe, Eric Brown, Marc Allard, and Lynn Bry. Diverse vectors and mechanisms spread new delhi metallo-β-lactamases among carbapenem-resistant enterobacteriaceae in the greater boston area. Antimicrobial agents and chemotherapy, 63(2):e02040–18, 2019.

[200] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. BMC bioinformatics, 10(1):1–9, 2009.

[201] Alessandra Carattoli, Ea Zankari, Aurora García-Fernández, Mette Voldby Larsen, Ole Lund, Laura Villa, Frank Møller Aarestrup, and Henrik Hasman. In silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrobial agents and chemotherapy, 58(7):3895–3903, 2014.

[202] Xiaobin Li, Yingzhou Xie, Meng Liu, Cui Tai, Jingyong Sun, Zixin Deng, and Hong-Yu Ou. oritfinder: a web-based tool for the identification of origin of transfers in dna sequences of bacterial mobile genetic elements. Nucleic acids research, 46(W1):W229–W234, 2018.

[203] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11):1422–1423, 2009.

[204] Sergey Nurk, Anton Bankevich, Dmitry Antipov, Alexey A Gurevich, Anton Korobeynikov, Alla Lapidus, Andrey D Prjibelski, Alexey Pyshkin, Alexander Sirotkin, Yakov Sirotkin, et al. Assembling single-cell genomes and mini-metagenomes from chimeric mda products. Journal of Computational Biology, 20(10):714–737, 2013.

[205] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology, 15(3):1–12, 2014.

[206] Aaron Petkau, Matthew Stuart-Edwards, Paul Stothard, and Gary Van Domselaar. Interactive microbial genome visualization with gview. Bioinformatics, 26(24):3125–3126, 2010.

[207] Patricia Siguier, Jocelyne Pérochon, L Lestrade, Jacques Mahillon, and Michael Chandler. Isfinder: the reference centre for bacterial insertion sequences. Nucleic acids research, 34(suppl_1):D32–D36, 2006.

[208] Ivica Letunic and Peer Bork. Interactive tree of life (itol) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic acids research, 44(W1):W242–W245, 2016.

[209] Ilia Minkin and Paul Medvedev. Scalable multiple whole-genome alignment and locally collinear block construction with sibeliaz. Nature communications, 11(1):1–11, 2020.

[210] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. Fast and accurate de novo genome assembly from long uncorrected reads. Genome research, 27(5):737–746, 2017.

[211] Ronan M Doyle, Denise M O'Sullivan, Sean D Aller, Sebastian Bruchmann, Taane Clark, Andreu Coello Pelegrin, Martin Cormican, Ernest Diez Benavente, Matthew J Ellington, Elaine McGrath, et al. Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: An inter-laboratory study. Microbial genomics, 6 (2), 2020.

[212] Xiaofang Jiang, A Brantley Hall, Timothy D Arthur, Damian R Plichta, Christian T Covington, Mathilde Poyet, Jessica Crothers, Peter L Moses, Andrew C Tolonen, Hera Vlamakis, et al. Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. Science, 363 (6423):181–187, 2019.

[213] CDC. Performance Standards for Antimicrobial Susceptibility Testing. 2017.

[214] Jianfeng Wang, Zhihui Zhou, Fang He, Zhi Ruan, Yan Jiang, Xiaoting Hua, and Yunsong Yu. The role of the type vi secretion system vgrg gene in the virulence and antimicrobial resistance of acinetobacter baumannii atcc 19606. PLoS One, 13(2):e0192288, 2018.

[215] Nachiket P Marathe, Fanny Berglund, Mohammad Razavi, Chandan Pal, Johannes Dröge, Sharvari Samant, Erik Kristiansson, and DG Joakim Larsson. Sewage effluent from an indian hospital harbors novel carbapenemases and integron-borne antibiotic resistance genes. Microbiome, 7(1):97, 2019.

[216] Brenda M Ryan, Thomas J Dougherty, Danielle Beaulieu, Julia Chuang, Brian A Dougherty, and John F Barrett. Efflux in bacteria: what do we really know about it? Expert opinion on investigational drugs, 10(8):1409–1422, 2001.

[217] Anaïs Le Rhun, Andrés Escalera-Maurer, Majda Bratovič, and Emmanuelle Charpentier. Crispr-cas in streptococcus pyogenes. RNA Biology, 16(4):380–389, 2019. doi: 10.1080/15476286.2019.1582974. URL https://doi.org/10.1080/15476286.2019.1582974. PMID: 30856357.

[218] Izabela Sitkiewicz, Nicole M. Green, Nina Guo, Ann M. Bongiovanni, Steven S. Witkin, and James M. Musser. Adaptation of group a streptococcus to human amniotic fluid. PLOS ONE, 5(3):1–13, 03 2010. doi: 10.1371/journal.pone.0009785. URL https://doi.org/10.1371/journal.pone.0009785.

[219] Máire Begley, Paul D. Cotter, Colin Hill, and R. Paul Ross. Identification of a novel two-peptide lantibiotic, lichenicidin, following rational genome mining for lanm proteins. Applied and Environmental Microbiology, 75(17):5451–5460, 2009. ISSN 0099-2240. doi: 10.1128/AEM.00730-09. URL https://aem.asm.org/content/75/17/5451.

[220] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

[221] Centers for Disease Control, Prevention (CDC, et al. Establishment of a national surveillance program for antimicrobial resistance in salmonella. MMWR. Morbidity and mortality weekly report, 45(5):110–111, 1996.

[222] L Tollefson. Fda reveals plans for antimicrobial susceptibility monitoring. Journal of the American Veterinary Medical Association, 208(4):459–460, 1996.

[223] Arthur W Pightling, Nicholas Petronella, and Franco Pagotto. The listeria monocytogenes core-genome sequence typer (lmcgst): a bioinformatic pipeline for molecular characterization with next-generation sequence data. BMC microbiology, 15(1):1–12, 2015.

[224] James B Pettengill, Arthur W Pightling, Joseph D Baugher, Hugh Rand, and Errol Strain. Real-time pathogen detection in the era of whole-genome sequencing and big data: comparison of k-mer and site-based methods for inferring the genetic distances among tens of thousands of salmonella samples. PLoS One, 11(11):e0166162, 2016.

[225] Werner Ruppitsch, Ariane Pietzka, Karola Prior, Stefan Bletz, Haizpea Lasa Fernandez, Franz Allerberger, Dag Harmsen, and Alexander Mellmann. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of listeria monocytogenes. Journal of clinical microbiology, 53(9):2869–2876, 2015.

[226] Alexander Mellmann, Stefan Bletz, Thomas Böking, Frank Kipp, Karsten Becker, Anja Schultes, Karola Prior, and Dag Harmsen. Real-time genome sequencing of resistant bacteria provides precision infection control in an institutional setting. Journal of clinical microbiology, 54(12):2874–2881, 2016.

[227] Nabil-Fareed Alikhan, Zhemin Zhou, Martin J Sergeant, and Mark Achtman. A genomic overview of the population structure of salmonella. PLoS genetics, 14(4):e1007261, 2018.

[228] Madison E Pearce, Nabil-Fareed Alikhan, Timothy J Dallman, Zhemin Zhou, Kathie Grant, and Martin CJ Maiden. Comparative analysis of core genome mlst and snp typing within a european salmonella serovar enteritidis outbreak. International journal of food microbiology, 274:1–11, 2018.

[229] Melanie Schirmer, Rosalinda D'Amore, Umer Z Ijaz, Neil Hall, and Christopher Quince. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC bioinformatics, 17(1):1–15, 2016.

[230] Luotong Wang, Li Qu, Longshu Yang, Yiying Wang, and Huaiqiu Zhu. Nanoreviser: an error-correction tool for nanopore sequencing based on a deep learning algorithm. Frontiers in Genetics, page 900, 2020.

[231] Juliane C Dohm, Philipp Peters, Nancy Stralis-Pavese, and Heinz Himmelbauer. Benchmarking of long-read correction methods. NAR Genomics and Bioinformatics, 2(2):lqaa037, 2020.

[232] Jamie K Lemon, Pavel P Khil, Karen M Frank, and John P Dekker. Rapid nanopore sequencing of plasmids and resistance gene detection in clinical isolates. Journal of clinical microbiology, 55(12):3530–3543, 2017.

[233] Zhao Chen, Dai Kuang, Xuebin Xu, Narjol Gonzalez-Escalona, David L Erickson, Eric Brown, and Jianghong Meng. Genomic analyses of multidrug-resistant salmonella indiana, typhimurium, and enteritidis isolates using minion and miseq sequencing technologies. PloS one, 15(7):e0235641, 2020.

[234] Narjol Gonzalez-Escalona, Marc A Allard, Eric W Brown, Shashi Sharma, and Maria Hoffmann. Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in shiga toxin-producing escherichia coli. PloS one, 14(7):e0220494, 2019.

[235] United States Food and Drug Administration. Bacteriological Analytical Manual. Aoac International, 1995.

[236] Andrea Ottesen, Padmini Ramachandran, Elizabeth Reed, James R White, Nur Hasan, Poorani Subramanian, Gina Ryan, Karen Jarvis, Christopher Grim, Ninalynn Daquiqan, et al. Enrichment dynamics of listeria monocytogenes and the associated microbiome from naturally contaminated ice cream linked to a listeriosis outbreak. BMC microbiology, 16(1):1–11, 2016.

[237] Kaire Loit, Kalev Adamson, Mohammad Bahram, Rasmus Puusepp, Sten Anslan, Riinu Kiiker, Rein Drenkhan, Leho Tedersoo, and Irina S. Druzhinina. Relative performance of minion (oxford nanopore technologies) versus sequel (pacific biosciences) third-generation sequencing instruments in identification of agricultural and forest fungal pathogens. Applied and Environmental Microbiology, 85(21):e01368–19, 2019. doi: 10.1128/AEM.01368-19. URL https://journals.asm.org/doi/abs/10.1128/AEM.01368-19.

[238] Lauren M. Petersen, Isabella W. Martin, Wayne E. Moschetti, Colleen M. Kershaw, Gregory J. Tsongalis, and Colleen Suzanne Kraft. Third-generation sequencing in the clinical laboratory: Exploring the advantages and challenges of nanopore sequencing. Journal of Clinical Microbiology, 58(1):e01315–19, 2019. doi: 10.1128/JCM.01315-19. URL https://journals.asm.org/doi/abs/10.1128/JCM.01315-19.

[239] Steve Hamner, Bonnie L. Brown, Nur A. Hasan, Michael J. Franklin, John Doyle, Margaret J. Eggers, Rita R. Colwell, and Timothy E. Ford. Metagenomic profiling of microbial pathogens in the little bighorn river, montana. International Journal of Environmental Research and Public Health, 16(7), 2019. ISSN 1660-4601. doi: 10.3390/ijerph16071097. URL https://www.mdpi.com/1660-4601/16/7/1097.

[240] Adriel Latorre-Pérez, Pascual Villalba-Bermell, Javier Pascual, and Cristina Vilanova. Assembly methods for nanopore-based metagenomic sequencing: a comparative study. Scientific reports, 10(1):1–14, 2020.

[241] James B Pettengill, Eugene McAvoy, James R White, Marc Allard, Eric Brown, and Andrea Ottesen. Using metagenomic analyses to estimate the consequences of enrichment bias for pathogen detection. BMC research notes, 5(1):1–7, 2012.

[242] Tina S Lusk, Andrea R Ottesen, James R White, Marc W Allard, Eric W Brown, and Julie A Kase. Characterization of microflora in latin-style cheeses by next-generation sequencing technology. BMC microbiology, 12(1):1–10, 2012.

[243] Karen G Jarvis, James R White, Christopher J Grim, Laura Ewing, Andrea R Ottesen, Junia Jean-Gilles Beaubrun, James B Pettengill, Eric Brown, and Darcy E Hanes. Cilantro microbiome before and after nonselective preenrichment for salmonella using 16s rrna and metagenomic sequencing. BMC microbiology, 15(1):1–13, 2015.

[244] Tina Lusk Pfefer, Padmini Ramachandran, Elizabeth Reed, Julie A. Kase, and Andrea Ottesen. Metagenomic description of preenrichment and postenrichment of recalled chapati atta flour using a shotgun sequencing approach. Genome Announcements, 6(21):e00305–18, 2018. doi: 10.1128/genomeA.00305-18. URL https://journals.asm.org/doi/abs/10.1128/genomeA.00305-18.

[245] Padmini Ramachandran, Elizabeth Reed, and Andrea Ottesen. Exploring the microbiome of callinectes sapidus (maryland blue crab). Genome Announcements, 6(22):e00466–18, 2018. doi: 10.1128/genomeA.00466-18. URL https://journals.asm.org/doi/abs/10.1128/genomeA.00466-18.

[246] Sylvia Ossai, Padmini Ramachandran, Andrea Ottesen, Elizabeth Reed, Angelo DePaola, and Salina Parveen. Microbiomes of american oysters (crassostrea virginica) harvested from two sites in the chesapeake bay. Genome Announcements, 5(30):e00729–17, 2017. doi: 10.1128/genomeA.00729-17. URL https://journals.asm.org/doi/abs/10.1128/genomeA.00729-17.

[247] Padmini Ramachandran, Elizabeth Reed, Seth Commichaux, Errol Strain, Angelo Depaola, Scott Rikard, and Andrea Ottesen. Characterization of the microbiota of oyster larvae (crassostrea virginica) and tank water from an aquaculture system with high and low larval survival rates. Genome Announcements, 6(25):e00597–18, 2018. doi: 10.1128/genomeA.00597-18. URL https://journals.asm.org/doi/abs/10.1128/genomeA.00597-18.

[248] Anna Townsend, Shaoting Li, David A. Mann, and Xiangyu Deng. A quasimetagenomics method for concerted detection and subtyping of salmonella enterica and e. coli o157:h7 from romaine lettuce. Food Microbiology, 92:103575, 2020. ISSN 0740-0020. doi: https://doi.org/10.1016/j.fm.2020.103575. URL https://www.sciencedirect.com/science/article/pii/S0740002020301647.

[249] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. Bioinformatics, 31(10):1674–1676, 2015.

[250] Mikhail Kolmogorov, Derek M Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy PL Smith, et al. metaflye: scalable long-read metagenome assembly using repeat graphs. Nature Methods, 17(11):1103–1110, 2020.

[251] Denis Bertrand, Jim Shaw, Manesh Kalathiyappan, Amanda Hui Qi Ng, M Senthil Kumar, Chenhao Li, Mirta Dvornicic, Janja Paliska Soldo, Jia Yu Koh, Chengxuan Tong, et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nature biotechnology, 37(8):937–944, 2019.

[252] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research, 27(5):722–736, 2017.

[253] Jue Ruan and Heng Li. Fast and accurate long-read assembly with wtdbg2. Nature methods, 17(2):155–158, 2020.

[254] Dmitry Antipov, Anton Korobeynikov, Jeffrey S McLean, and Pavel A Pevzner. hybridspades: an algorithm for hybrid assembly of short and long reads. Bioinformatics, 32(7):1009–1015, 2016.

[255] Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one, 9 (11):e112963, 2014.

[256] René L Warren, Lauren Coombe, Hamid Mohamadi, Jessica Zhang, Barry Jaquish, Nathalie Isabel, Steven JM Jones, Jean Bousquet, Joerg Bohlmann, and Inanç Birol. ntedit: scalable genome sequence polishing. Bioinformatics, 35(21):4430–4432, 2019.

[257] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. Fast and accurate de novo genome assembly from long uncorrected reads. Genome research, 27(5):737–746, 2017.

[258] Yi Chen, Yan Luo, Phillip Curry, Ruth Timme, David Melka, Matthew Doyle, Mickey Parish, Thomas S Hammack, Marc W Allard, Eric W Brown, et al. Assessing the genome level diversity of listeria monocytogenes from contaminated ice cream and environmental samples linked to a listeriosis outbreak in the united states. PLoS One, 12(2):e0171389, 2017.

[259] Richard J Roberts, Mauricio O Carneiro, and Michael C Schatz. The advantages of smrt sequencing. Genome biology, 14(6):1–4, 2013.

[260] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using minhash. Genome biology, 17(1):1–14, 2016.

[261] Karin Lagesen, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Stærfeldt, Torbjørn Rognes, and David W Ussery. Rnammer: consistent and rapid annotation of ribosomal rna genes. Nucleic acids research, 35(9):3100–3108, 2007.

[262] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. Nature methods, 9(4):357–359, 2012.

[263] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. Bioinformatics, 25(16):2078–2079, 2009.

[264] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18):3094–3100, 2018.

[265] Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome research, 18(11): 1851–1858, 2008.

[266] Kiran Javkar, Hirak Sarkar, Hugh Rand, Rob Patro, and Mihai Pop. Simile: Discover similer genomic regions shared across a collection of metagenomic samples. `https://github.com/KiranJavkar/SIMILE`, 2022.

[267] Alison R Erickson, Brandi L Cantarel, Regina Lamendella, Youssef Darzi, Emmanuel F Mongodin, Chongle Pan, Manesh Shah, Jonas Halfvarson, Curt Tysk, Bernard Henrissat, et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of crohn's disease. PloS one, 7(11): e49138, 2012.

[268] Human Microbiome Project Consortium et al. Structure, function and diversity of the healthy human microbiome. nature, 486(7402):207, 2012.

[269] Don A Cowan, Jean-Baptiste Ramond, Thulani P Makhalanyane, and Pieter De Maayer. Metagenomics of extreme environments. Current opinion in microbiology, 25:97–102, 2015.

[270] Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Shotgun metagenomics, from sampling to analysis. Nature biotechnology, 35(9):833–844, 2017.

[271] David Zeevi, Tal Korem, Anastasia Godneva, Noam Bar, Alexander Kurilshikov, Maya Lotan-Pompan, Adina Weinberger, Jingyuan Fu, Cisca Wijmenga, Alexandra Zhernakova, et al. Structural variation in the gut microbiome associates with host health. Nature, 568(7750):43–48, 2019.

[272] Dongwan D Kang, Jeff Froula, Rob Egan, and Zhong Wang. Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ, 3:e1165, 2015.

[273] Helle Krogh Pedersen, Valborg Gudmundsdottir, Henrik Bjørn Nielsen, Tuulia Hyotylainen, Trine Nielsen, Benjamin AH Jensen, Kristoffer Forslund, Falk Hildebrand, Edi Prifti, Gwen Falony, et al. Human gut microbes impact host serum metabolome and insulin sensitivity. Nature, 535(7612):376–381, 2016.

[274] Naseer Sangwan, Fangfang Xia, and Jack A Gilbert. Recovering complete and draft population genomes from metagenome datasets. Microbiome, 4(1):1–11, 2016.

[275] Xiaofang Jiang, Andrew Brantley Hall, Ramnik J Xavier, and Eric J Alm. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. PloS one, 14(12): e0223680, 2019.

[276] Sumaiya Nazeen, Yun William Yu, and Bonnie Berger. Carnelian uncovers hidden functional patterns across diverse study populations from whole metagenome sequencing reads. Genome biology, 21(1):1–18, 2020.

[277] Brian D Ondov, Gabriel J Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B Buck, and Adam M Phillippy. Mash screen: high-throughput sequence containment estimation for genome discovery. Genome biology, 20(1):1–13, 2019.

[278] Huan Fan, Anthony R Ives, Yann Surget-Groba, and Charles H Cannon. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC genomics, 16(1):1–18, 2015.

[279] Boopathy Usharani. Metagenomics study of the microbes in constructed wetland system treating sewage. International Letters of Natural Sciences, 74, 2019.

[280] Ohana YA Costa, Mattias De Hollander, Agata Pijl, Binbin Liu, and Eiko E Kuramae. Cultivation-independent and cultivation-dependent metagenomes reveal genetic and enzymatic potential of microbial community involved in the degradation of a complex microbial polymer. Microbiome, 8(1):1–19, 2020.

[281] Valeria Sagheddu, Vania Patrone, Francesco Miragoli, Edoardo Puglisi, and Lorenzo Morelli. Infant early gut colonization by lachnospiraceae: high frequency of ruminococcus gnavus. Frontiers in pediatrics, 4:57, 2016.

[282] Mirco Vacca, Giuseppe Celano, Francesco Maria Calabrese, Piero Portincasa, Marco Gobbetti, and Maria De Angelis. The controversial role of human gut lachnospiraceae. Microorganisms, 8(4):573, 2020.

[283] Nicolai Karcher, Edoardo Pasolli, Francesco Asnicar, Kun D Huang, Adrian Tett, Serena Manara, Federica Armanini, Debbie Bain, Sylvia H Duncan, Petra Louis, et al. Analysis of 1321 eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. Genome biology, 21:1–27, 2020.

[284] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28(23):3150–3152, 2012.

[285] Rebecca A Gladstone, Alan McNally, Anna K Pöntinen, Gerry Tonkin-Hill, John A Lees, Kusti Skytén, François Cléon, Martin OK Christensen, Bjørg C Haldorsen, Kristina K Bye, et al. Emergence and dissemination of antimicrobial resistance in escherichia coli causing bloodstream infections in norway in 2002–17: a nationwide, longitudinal, microbial population genomic study. The Lancet Microbe, 2(7):e331–e341, 2021.

[286] Souvik Bhattacharyya, David M Walker, and Rasika M Harshey. Dead cells release a 'necrosignal'that activates antibiotic survival pathways in bacterial swarms. Nature communications, 11(1):1–12, 2020.

[287] Jerónimo Rodríguez-Beltrán, Javier DelaFuente, Ricardo Leon-Sampedro, R Craig MacLean, and Alvaro San Millan. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nature Reviews Microbiology, 19(6):347–359, 2021.

[288] Ellie Harrison and Michael A Brockhurst. Plasmid-mediated horizontal gene transfer is a coevolutionary process. Trends in microbiology, 20(6):262–267, 2012.

[289] Derek M Bickhart, Mikhail Kolmogorov, Elizabeth Tseng, Daniel M Portik, Anton Korobeynikov, Ivan Tolstoganov, Gherman Uritskiy, Ivan Liachko, Shawn T Sullivan, Sung Bong Shin, et al. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. Nature biotechnology, 40(5):711–719, 2022.

[290] Eitan Yaffe and David A Relman. Tracking microbial evolution in the human gut using hi-c reveals extensive horizontal gene transfer, persistence and adaptation. Nature microbiology, 5(2):343–353, 2020.

[291] Suraj Nair, Kiran Javkar, Jiahui Wu, and Vanessa Frias-Martinez. Understanding cycling trip purpose and route choice using gps traces and open data. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 3(1), mar 2019. doi: 10.1145/3314407. URL https://doi.org/10.1145/3314407.

[292] Debjani Saha, Anna Chan, Brook Stacy, Kiran Javkar, Sushant Patkar, and Michelle L. Mazurek. User attitudes on direct-to-consumer genetic testing. In 2020 IEEE European Symposium on Security and Privacy (EuroS&P), pages 120–138, 2020. doi: 10.1109/EuroSP48549.2020.00016.

[293] Elisha M Wood-Charlson, Deanna Auberry, Hannah Blanco, Mark I Borkum, Yuri E Corilo, Karen W Davenport, Shweta Deshpande, Ranjeet Devarakonda, Meghan Drake, William D Duncan, et al. The national microbiome data collaborative: enabling microbiome science. Nature Reviews Microbiology, 18 (6):313–314, 2020.

[294] Nehal Mohamed Ali, Marwa Mostafa Abd El Hamid, and Aliaa Youssif. Sentiment analysis for movies reviews dataset using deep learning models. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 9, 2019.

[295] Ashwin Karale. The challenges of iot addressing security, ethics, privacy, and laws. Internet of Things, 15:100420, 2021.

[296] Sarah Fendrick. The role of privacy law in genetic research. ISJLP, 4:803, 2008.

[297] Andelka M Phillips. Think before you click: ordering a genetic test online. GPSolo, 32:72, 2015.

[298] Jian-Yu Jiao, Lan Liu, Zheng-Shuang Hua, Bao-Zhu Fang, En-Min Zhou, Nimaichand Salam, Brian P Hedlund, and Wen-Jun Li. Microbial dark matter coming to light: challenges and opportunities. National Science Review, 8(3): nwaa280, 2021.

[299] Joan Martí-Carreras, Alejandro Rafael Gener, Sierra D Miller, Anderson F Brito, Christiam E Camacho, Ryan Connor, Ward Deboutte, Cody Glickman, David M Kristensen, Wynn K Meyer, et al. Ncbi's virus discovery codeathon: Building "five"—the federated index of viral experiments api index. Viruses, 12(12):1424, 2020.