ABSTRACT

Title of Dissertation: METHOD VALIDATION AND DEVELOPMENT FOR THE METAGENOMIC EXPLORATION OF MICROBIAL COMMUNITIES

Seth Commichaux
Doctor of Philosophy, 2022

Dissertation directed by: Professor Mihai Pop
Department of Computer Science

Dr. Hugh Rand
Food and Drug Administration

Our world is inhabited and shaped by diverse and complex microbial communities which we are only beginning to characterize and understand. With the advent of affordable high-throughput sequencing, the study of the genomic content of microbial communities, metagenomics, has accelerated our understanding of their impact on human and environmental health. The increasing number of datasets produced by metagenomic studies provide many opportunities for novel bioinformatic analyses and for the development of computational methods. However, careful benchmarking and validation are also important undertakings to ensure the integrity of methods and research in such a rapidly developing field. Here, we explored several problems in metagenomics by benchmarking existing methods and technologies, developing new methods, recommending best practices, and highlighting opportunities for future work.

First, microbial gene catalogs document and organize the genes found in microbial communities and provide a reference for the standardized analysis of metagenomic data. Although commonly used to explore the intersection between microbiomes, humans, and ecosystems, the methods used for their construction and effectiveness for metagenomic analyses had not been critically evaluated. Our analysis highlighted important limitations of gene catalogs, opportunities for future research, and allowed us to recommend best practices.

Second, we assessed if nanopore long read sequencing could expedite the accurate reconstruction of a pathogen genome from a microbial community. The investigation of foodborne illness outbreaks routinely uses short-read whole genome sequencing of pure culture pathogen colonies. However, culturing is a bottleneck and short reads cannot span all bacterial genomic repeats, often leading to fragmented assemblies. Our results showed that the integration of long-read sequencing could expedite the public health response by reconstructing complete pathogen genomes from a microbial community after limited culturing. Additionally, our evaluation of state-of-the-art assembly tools identified biases and areas for improvement.

Third, we describe taxaTarget, a supervised learning approach for the taxonomic classification of microeukaryotes in metagenomic data. Metagenomics has been underutilized for microeukaryotes due to the many computational challenges they present. Existing tools often implement universal sequence similarity cutoffs which ignore that sequences can evolve at different rates and, thus, have different discriminatory power. We show that a data-driven approach to determining classification thresholds can result in higher sensitivity and precision than existing tools.

Fourth, we explored the use of horizontally transferred plasmids to relate an outbreak strain to the microbiome of a suspected environmental source. The investigation of the 2020 red onion outbreak recovered the outbreak strain from patients but not the farms implicated as the likely source of contamination. Our analysis identified highly similar plasmids in the outbreak strain and environmental isolates collected from the farms, which supported a connection between the outbreak strain and the implicated farms. Additionally, we highlighted the need for more detailed and accurate metadata, more extensive environmental sampling, and a better understanding of plasmid molecular evolution before such analyses can be added to the public health response.

METHOD VALIDATION AND DEVELOPMENT FOR THE METAGENOMIC EXPLORATION OF MICROBIAL COMMUNITIES

by

Seth Commichaux

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/Advisor
Dr. Hugh Rand, Co-Advisor
Professor Michael Cummings, Dean’s Representative
Professor Najib El-Sayed,
Dr. Andrea Ottesen

Preface

The algorithms, software, and results in this dissertation have either been published in peer-reviewed journals, have been submitted and are under review, or are currently under preparation for submission. At the time of this writing, Chapters 2, and 3 have already been published, Chapter 4 has been submitted and is 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 co-authors whose dedication and knowledge resulted in greatly improved works.

- Chapter 2
    - Commichaux, Seth, Nidhi Shah, Jay Ghurye, Alexander Stoppel, Jessica A. Goodheart, Guillermo G. Luque, Michael P. Cummings, and Mihai Pop. "A critical assessment of gene catalogs for metagenomic analysis." *Bioinformatics* 37, no. 18 (2021). 2848-2857. This work was co-first-authored with Nidhi Shah. My contributions include (1) reviewing the literature, (2) experimental design, (3) running experiments, (4) data analysis, and (5) writing the manuscript.
- Chapter 3
    - Commichaux, Seth, Kiran Javkar, Padmini Ramachandran, Niranjan Nagarajan, Denis Bertrand, Yi Chen, Elizabeth Reed et al. "Evaluating the accuracy of *Listeria monocytogenes* assemblies from quasimetagenomic samples using long and short reads." *BMC genomics* 22, no. 1 (2021): 1-18. This work was co-first-authored with Kiran Javkar. My contributions include (1) reviewing the literature,

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

- Chapter 4
    - Commichaux, Seth, Kiran Javkar, Harihara Muralidharan, Padmini Ramachandran, Andrea Ottesen, and Mihai Pop. "taxaTarget: Fast, sensitive, and precise classification of microeukaryotes in metagenomic data." (2022). Preprint on Research Square (DOI: https://doi.org/10.21203/rs.3.rs-1186624/v2). My contributions include (1) reviewing the literature, (2) experimental design, (3) running experiments, (4) data analysis, (5) design and implementation of the algorithms, (6) updating and maintaining the software, and (7) writing the manuscript.
- Chapter 5
    - Commichaux, Seth, 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." (2022). To be submitted to BMC Genomics. My contributions include (1) reviewing the literature, (2) experimental design, (3) running experiments, (4) data analysis, and (5) writing the manuscript.

Other publications to which I have contributed during my PhD.

- Ramachandran, Padmini, 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, no. 25 (2018): e00597-18.

- Cagle, Robin, Padmini Ramachandran, Elizabeth Reed, Seth Commichaux, Mark K. Mammel, David W. Lacher, Daniela Miller et al. "Microbiota of the hickey run tributary of the Anacostia River." *Microbiology resource announcements* 8, no. 12 (2019): e00123-19.
- Sakowski, Eric, Gherman Uritskiy, Rachel Cooper, Maya Gomes, Michael R. McLaren, Jacquelyn S. Meisel, Rebecca L. Mickol et al. "Current state of and future opportunities for prediction in microbiome research: report from the Mid-Atlantic Microbiome Meet-up in Baltimore on 9 January 2019." *Msystems* 4, no. 5 (2019): e00392-19.
- Reed, Elizabeth, Robert Sanderson, Barbara Sanderson, Padmini Ramachandran, Seth Commichaux, Eric Brown, Jie Zheng, Errol Strain, and Andrea Ottesen. "Water Microbiota from a Commercial Alfalfa Sprout Crop over 4 Days of Growth." *Microbiology Resource Announcements* 10, no. 2 (2021): e00906-20.
- Ottesen, Andrea, Brandon Kocurek, Padmini Ramachandran, Elizabeth Reed, Seth Commichaux, Gunnar Engelbach, Mark Mammel et al. "Advancing antimicrobial resistance monitoring in surface waters with metagenomic and quasimetagenomic methods." *bioRxiv* (2022).
- Lafleur, John, Jacquelyn S. Meisel, Seth Commichaux, Richard L. Amdur, Mihai Pop, and Mark W. Silby. "Paraoxonase and acylated homoserine lactones in urine from patients with urinary tract infections--relationship to microbial diversity by 16S rRNA gene sequencing." *bioRxiv* (2021).

Dedication

To my family, friends, mentors, teachers, facilitators, and collaborators for all their support and wisdom.

Acknowledgements

I have had the chance to meet and be supported by so many wonderful people leading up to and during the PhD program. I cannot remember all their names nor all their contributions to my growth and happiness, but I am truly grateful and hope I can pay forward a fraction of what I have received.

First, I sincerely appreciate my advisors—Mihai Pop and Hugh Rand—for being wonderful and patient mentors, for training me, allowing me to pursue my research interests, and helping me transform those interests into published works. The many conversations I have had through the highs and the lows were all meaningful and important to my continued growth as a researcher and a person.

I also want to thank my committee members. Andrea Ottesen for being a source of creative energy, sound advice, community, and scientific curiosity. I cannot fathom myself ever attaining your level of productivity, creativity, and encouragement! Michael Cummings for helping me improve my technical skills and to progress in scientific rigor and professionalism. It was wonderful to collaborate at the BEST summer school and on the resulting gene catalog project. Najib El-Sayed for providing mentorship early in the program when I was still trying to find my research direction. I also appreciate all the constructive feedback and the insightful questions that helped me grow on my research path.

I have also had many amazing lab mates who I could always rely upon for brainstorming, venting, collaborating, networking, and to have a good laugh— Mathieu Almeida, Domenick Braccia, Brian Brubach, Marcus Fedarko, Jay Ghurye, Kiran Javkar, Tu Luan, Shiva Mehravaran, Jacquelyn Stephanie Meisel, Hari Muralidharan, Dan Nasko, Nate Olson, Joseph

Paulson, Saul Sarria, Jeremy Selengut, Nidhi Shah, Brook Stacy, Alexander Stoppel, Dylan Taylor, and Todd Treangen.

It also would have been impossible to succeed without the CBCB, CBBG, BISI, and UMIACS communities—including fellow students, faculty, and staff.

I also have many people to thank at the FDA. It was through a fellowship offered by the FDA that I had the chance to collaborate with researchers there, eventually leading to a job as a bioinformatician. Special thanks to Jayanthi Gangiredla, Sara Handy, David Lacher, Yan Luo, Mark Mammel, Amit Mukherjee, James Pettengill, Arthur Pightling, Padmini Ramachandran, Liz Reed, Marianne Solomotis, Errol Strain, Amanda Windsor.

I would also like to thank the many mentors that encouraged and prepared me to pursue a PhD, both during my undergrad and master's programs: Mark Daniels, Nathan Edwards, Gamini Gunawardena, Merrill Halling, Haagen Klaus, Olga Kopp, Ruhul Kudus, Catherine Malone, James Price, and Karen Ross.

Lastly, I would like to thank all my family and friends for all their love and support. And a special thanks to my wife and partner in crime, Eden Mesfun. She has been my rock through all the storms, and I would not have made it to the finish line without her love.

Table of Contents

List of Tables

List of Figures

1. Introduction

Our world is inhabited and shaped by diverse and complex communities of microscopic organisms that have been evolving since life emerged on Earth over 3.5 billion years ago [1]. Microbial communities, collectively referred to as the microbiome, often consist of complex mixtures of bacteria, archaea, viruses, and eukaryotes. These microbes are incredibly adaptable and can be found in bodies of water (e.g., oceans, rivers, aquifers), the soil, the air, as well as living in and on plants and animals. Microbes can even be found in extreme environments such as circulating in the stratosphere, miles underground, within rocks and glacial ice, in acidic sulfur springs, and near hydrothermal vents [2-6]. Despite constituting the majority of the biomass on Earth, most microbial taxa and their ecological contributions remain uncharacterized [7]. Although the most well-known microbes are those capable of causing disease, the vast majority are beneficial and complement the functions of their hosts and ecosystems [8, 9].

For more than a century, the standard techniques for characterizing microorganisms relied on microscopic observation and enrichment culture. However, microbes are difficult to differentiate visually, and most cannot currently be cultured. With the advent of high throughput sequencing, the genomes of single organisms could be sequenced as well as mixtures of organisms directly from the environment. The direct sequencing and analysis of sequence data from mixtures of organisms is called metagenomics. Metagenomics has provided a powerful tool for surveying microbiomes across the world, revealing their genetic diversity and ecological contributions.

1.1 DNA sequencing

Initially, the study of microbes was limited to microscopic observation and culturing; however, most microbes are not easy to visually differentiate or to culture in laboratory environments [10]. DNA sequencing, the determination of the base pair composition of DNA strands, has allowed for a much more detailed study of the genetic diversity and functional potential of microbes. Further, it has led to an explosion in the number of sequenced genomes (all the DNA from an organism), sequenced metagenomes (all the DNA from an environmental sample), and classified microbial species [11, 12].

DNA sequencing typically begins by extracting the DNA from cells and fragmenting it. Fragmentation of the DNA is required because none of the currently available sequencing technologies can sequence all genomes in a single pass from beginning to end. The sequenced fragments output by the sequencing machines are referred to as reads.

The various DNA sequencing technologies developed since the 1970's can be broadly classified into three generations [13]. First-generation sequencing, commonly called Sanger sequencing, uses bacterial cloning, DNA polymerases, and modified nucleotides to terminate the elongation of oligonucleotides along a template DNA at different points. The oligonucleotides are then size separated by electrophoresis to determine the sequence. Although Sanger sequencing is still used for small scale sequencing projects and for determining highly repetitive regions of genomes, it has largely been superseded by second-generation sequencing which is more automatable and higher throughput [13].

Second-generation sequencing (provided by platforms such as the Illumina MiSeq) is often referred to as short read sequencing because the reads are typically 300 nucleotides long or shorter. For comparison, Sanger sequencing produces reads up to ~2,000 nucleotides long.

Whereas sequencing a single read with Sanger sequencing was relatively slow and labor intensive, platforms like the Illumina NovaSeq sequence millions of reads in parallel, producing hundreds of millions of reads in 24 hours. Further, short read sequencing attains high throughput while maintaining high-fidelity, with ~99.9% of the bases called accurately for Illumina platforms [13]. However, short read sequencing can be biased towards representing genomic regions with a GC-content between 45% to 65% [14].

Third-generation sequencing, often called single molecule or long read sequencing, allows for the sequencing of single DNA molecules without amplification, which is required by the first- and second-generation sequencing platforms. In addition to the determination of the base pair composition of DNA strings, some third-generation sequencing platforms can also read DNA base (epigenetic) modifications like methylation. In terms of throughput, long read sequencing platforms, like the Oxford Nanopore MinIon, may only produce tens or hundreds of thousands of reads per run, but usually with a mean length of thousands to tens of thousands of nucleotides, with some reads being hundreds of thousands or even millions of nucleotides long. A current disadvantage of long read sequencing is the high error rate for individual reads, 5% or higher, consisting of insertions, deletions, and miscalled bases [13]. Additionally, long read sequencing can be subject to compositional biases. For example, genomic regions with skewed GC-content deviating from 50% can be underrepresented in PacBio long read data [14] and reads with low GC-content tend to have less errors than high GC-content reads in nanopore long read data [15].

1.2 Sequencing data: Genomics, metagenomics, and quasimetagenomics

The genome of an organism is the complete set of DNA from that organism. Genomics is the sequencing and analysis of genomes, often focusing on their annotation, structure, function, and evolution. For genomics, selective culture enrichment (a technique for isolating pure colonies of a single organism from a microbial community) is often required to obtain enough DNA for sequencing. In contrast, the culture-independent sequencing and analysis of the total organismal DNA directly extracted from a sample is called metagenomics [10]. Metagenomics has revolutionized our understanding of microbiomes, allowing for the direct exploration of natural microbial communities and the reconstruction of genomes from microbes that are currently unculturable [16-18]. Quasimetagenomics, a middle ground between genomics and metagenomics, is the sequencing and analysis of microbial communities whose species diversity has been partially reduced by selective culture enrichment (not completely reduced to pure colonies)—typically for a targeted taxon [19]. Quasimetagenomics is often used to expedite the targeted recovery of pathogen genomes from background microbial communities [19-21].

1.2.1 Assembly of metagenomic sequencing data

Metagenomic reads can be thought of as randomly sampled subsequences from the genomes found in a microbial community. Computer algorithms can be used to assemble the reads into more contiguous sequences called contigs, sometimes reconstructing complete genomes. Assembling contigs is still a computationally challenging problem for metagenomics, because of repeated sequences within genomes, shared regions between genomes, uneven abundances of organisms, sequencing errors, and strain-level variation [22]. Recently, long read sequencing has greatly increased the contiguity of genomic assemblies, often resolving repeats that short read sequencing could not [23]. For metagenomics, there is currently a performance tradeoff between short and long read sequencing. Although short read sequencing is higher throughput with a lower sequencing error rate, the assemblies tend to be highly fragmented. In contrast, long read sequencing is lower throughput, potentially sampling fewer reads from each species in the metagenome, with a higher sequencing error rate, but provides more contiguous information [24].

1.2.2 The analysis of metagenomic sequencing data

Many different bioinformatic analyses can be performed to characterize metagenomic datasets [25]. Often, researchers want to know what species are in their samples, what functions the species are performing (e.g., how they obtain energy, if they can cause disease), and in what abundances. For example, to optimize the treatment of wastewater, metagenomics has been used to identify microbial species that efficiently decompose organic matter and remove pollutants, as well as to understand how the abundance of pathogens changes during treatment [26].

In metagenomics, the annotation of species in a sample and the functions they perform is called taxonomic classification and functional profiling, respectively. A common starting point for taxonomic and functional annotation is to align the metagenomic reads and/or contigs to a reference database of annotated sequences, such as the proteins in the UniRef [27] database or the genomes in the GenBank database [28]. Various alignment-based metrics of sequence similarity, such as percent identity or the bitscore, are then used to assign taxonomic and functional annotations to the reads/contigs. Further, the read alignments to the reference database or back to the contigs can be used to estimate the relative abundances of the taxa and functions in the sample. A limitation of this reference-based approach is that it is biased towards the identification of taxa and functions represented in the reference databases, potentially leaving many other taxa and functions undetected [25]. Nonetheless, taxonomic and functional profiles have been used to study and compare microbiomes for a wide range of applications such as exploring the tremendous functional diversity of proteins in the ocean [29] and metagenome-wide association studies (MWAS) to discover biomarkers in the human gut microbiome for Type II diabetes [30].

1.3 Microbial gene catalogs

Since the development of high throughput sequencing, hundreds of thousands of microbial genomes and metagenomes have been sequenced and assembled. Researchers are interested in efficiently documenting the sequence diversity that has been observed in specific microbiomes, like the human gut, to facilitate future studies. One strategy to address this challenge is the construction of microbial gene catalogs—data structures that document the genes, functions, and taxa found in microbial communities and that provide a non-redundant reference for the standardized analysis of microbes across metagenomic samples and studies.

Gene catalogs are typically constructed by clustering the genes found in genomic and metagenomic contigs, from a specific microbiome, into a non-redundant set of representative genes. Clustering is used to remove trivial differences between sequences due to fragmentary data, sequencing errors, or strain-level variations. Typically, fast heuristic clustering approaches like CD-Hit [31] are used to accomplish such large-scale clustering.

For analysis, gene catalogs from different microbiomes can be compared, for example to explore how microbial communities differ between animal hosts. Gene catalogs can also be used to analyze metagenomic data, where metagenomic reads are aligned to the taxonomically and functionally annotated genes of the catalog to identify the genes, functions, and species—as well as their abundances—in the samples. Such an approach can be used to understand how diet in different human populations affects the microbial composition of the gut.

The first study to explicitly create a gene catalog used it to reveal the tremendous diversity of bacterial functions in the ocean [29]. Using slightly different methods, the MetaHit consortium provided the first gene catalog for the human gut microbiome [32]. Since then, gene catalogs have become ubiquitous in the analysis of metagenomic datasets and have been created for the gut microbiota of multiple animals, ocean bacteria, soil bacteria, as well as the human vagina and respiratory tract.

1.4 Taxonomic classification

The identification of an organism's evolutionary lineage is referred to as taxonomic classification. The modern system of taxonomy was founded by Carl Linnaeus in the 18th century, who hierarchically classified organisms based upon shared physical traits. Since the advent of high throughput DNA sequencing, taxonomy has rapidly shifted towards a system that reflects our understanding of the evolutionary relationships between taxa using molecular information [33]. Such molecular-based methods have been able to distinguish many microbial species that were phenotypically indistinguishable, resulting in an explosion in the number of characterized microbial species. However, next generation sequencing has also highlighted the dynamism of microbial genome evolution and the difficulty of finding genomic features that define microbial species [34, 35].

The current taxonomy subdivides all cellular life into three domains—archaea, bacteria, and eukaryotes. The domains are further subdivided into six or seven main taxonomic levels—kingdom, phylum, class, order, family, genus, and species, where kingdom is usually not used for bacteria and archaea.

For metagenomics, taxonomic classification is the computational process of inferring which taxonomic ranks are present in a microbial community. Commonly, taxonomic classification is accomplished by searching against a taxonomically characterized database and using sequence similarity to estimate evolutionary relatedness. A fundamental goal for metagenomics is to be able comprehensively classify all the taxa in any sample. However, comprehensive taxonomic classification of metagenomes is currently not possible because most microbial taxa are still uncharacterized.

1.5 Molecular epidemiology

If hospitals across the United States start reporting cases of Salmonellosis, it is critical to quickly determine if the cases are connected to a common source of contamination (e.g., chicken thighs produced by a specific company) so that action can be taken to stop the outbreak and to minimize the number and severity of illnesses. The analysis of molecular data, such as the *Salmonella* genomes isolated from sick patients, can provide valuable information about the potential source(s) of contamination, the geographic location it originated from, and if there are virulence factors (e.g., antimicrobial resistance genes) that need to be considered for effective medical treatment.

Molecular epidemiology refers to the use of molecular methods to monitor illness outbreaks, infection chains, and patterns of disease transmission in populations. Molecular epidemiology also seeks to identify the factors that determine the transmission, manifestation, and progression of infectious diseases. The unit being tracked could be a taxon (e.g., species or strain) or mobile genetic element (e.g., genome islands, plasmids), or even virulence or antimicrobial resistance factors. Regardless of the unit, a molecular typing scheme must be developed that provides sufficient discriminatory power to assess the presence and sequence diversity of the target pathogen [36].

Before molecular methods (e.g., PCR, high throughput sequencing), the identification and classification of bacterial pathogens was mainly accomplished by phenotypic assays (e.g., biochemical assays, enzymatic assays, antimicrobial susceptibility testing, serotyping). However, DNA sequencing helped show that some phenotypic characteristics were driven by complex genetic factors which did not strictly correlate with taxonomy (e.g., antibiotic resistance carried by plasmids) [36].

The development of molecular methods for the analysis of illness outbreaks can be broadly classified into four generations. The first generation was the analysis of isolate plasmid content by agarose-gel electrophoresis. However, plasmid profiles are not stable across time and are thus not reliable for typing. The second generation involved the comparative analysis of chromosome fragmentation by restriction enzymes using agarose-gel electrophoresis. This approach was limited to the analysis of relatively short DNA fragments (~50,000 bp or less) and restriction enzymes that recognized common motifs, which made accurate discrimination difficult. These problems were solved by the third generation pulsed field gel electrophoresis (PFGE), which used a more dilute agarose and alternating electrical fields to separate DNA fragments up to 10 million bp long [37].

Although PFGE is still the “gold standard” for many pathogens in clinical settings, whole genome sequencing (WGS) has gained ground as the preferred method [36]. WGS allows for high resolution differentiation of genomes based on single nucleotide polymorphisms (SNPs), allelic differences in genes, and structural variations (e.g., insertions, deletions, recombination events) [36]. WGS also allows for the identification and comparison of virulence and antimicrobial resistance factors, mobile elements, and (with the use of long-read sequencing platforms like PacBio and Oxford Nanopore) epigenetic modifications. To assist with WGS-based efforts for molecular epidemiology, large databases of pathogen genomes have been collected and curated such as GenomeTrakr [38].

One drawback of whole genome sequencing is the requirement of selective culture enrichment to isolate the pathogen from its microbial background for DNA sequencing. Ideally, culture independent methods, like metagenomics, would be used directly to learn about the pathogen and its microbial community. However, metagenomics has seen limited success for pathogen detection because pathogens often occur at low abundance within microbial communities, making the cost of attaining the necessary sequencing coverage to reconstruct their genomes prohibitive. As a middle ground, quasimetagenomics has seen success for reducing culturing times, recovering pathogen genomes from reduced diversity microbial communities, and understanding the microbiomes that pathogens come from [20, 21, 39].

1.6 Method development and validation for metagenomics

The field of metagenomics is rapidly evolving with an ever-increasing number of sequencing datasets, studies, applications, advances in sequencing technologies, and new computational tools [25]. Despite these developments, much of the microbial universe is still “dark matter”, full of microbes that cannot yet be cultured, with uncharacterized genetics, taxonomy, functions, and ecological interactions; there are also a plethora of uncharacterized genetic elements (e.g., plasmids, viruses, viroids, transposons) [17]. Further, metagenomic data analysis presents many computational challenges such as sequence clustering, annotation, organization, and phylogenetics, as well as genome assembly and comparison.

For such a rapidly developing field with so many unknowns, careful benchmarking and validation are important undertakings to ensure the integrity of methods and research. Benchmarking and validation are especially imperative because of the many applications of metagenomics, often right at the intersection of microbiomes and the health of humans and ecosystems. For example, metagenomics has been used to verify the microbial composition of probiotics by regulatory agencies like the Food and Drug Administration (FDA) [40]. For such applications, it is critically important that all the species in the probiotic are correctly identified to avoid missing or falsely identifying pathogens (potentially leading to incorrect regulatory action against the manufacturer) and to verify that the microbial composition matches the label of the product.

1.7 Contributions

The metagenomic study of microbiomes presents many bioinformatic challenges. For example, how to document and organize the sequence diversity observed in the sometimes thousands of metagenomic datasets from a specific microbiome e.g., the human gut, to provide a reference for the community. Another challenge is the comprehensive taxonomic classification of all sequences in a metagenomic sample. Currently, there are far more uncharacterized than characterized microbial species and there is a bias towards the study of bacteria compared to the other domains of life. There are also many bioinformatic challenges associated with the many applications of metagenomics such as the characterization of pathogens in microbial communities. For example, pathogens often occur at low abundance in complex microbial communities making it difficult to accurately reconstruct their genomes from metagenomes. Additionally, during food poisoning outbreaks, pathogens can be exposed to many environments in a short timeframe (e.g., irrigation water, farm soil, a growing onion, a processing plant, packaging, guacamole, the human gut) leading to rapid genomic evolution through horizontal transfer with local microbial communities, often having an unknown impact on their virulence to humans. Motivated by these bioinformatic challenges, we made the following contributions which are briefly summarized below and described in greater detail in the following chapters.

In Chapter 2, we assess the construction of microbial gene catalogs and their use for metagenomic analyses. Although gene catalogs are commonly used, they have not been critically evaluated for their effectiveness as a basis for metagenomic analyses. As a case study, we investigate one such catalog, the Integrated Gene Catalog (IGC); however, our observations apply broadly to most gene catalogs constructed to date. We focus on both the approach used to construct the IGC and on its effectiveness when used as a reference for microbiome studies. Our results highlight important limitations of the approach used to construct the IGC and call into question the broad usefulness of gene catalogs more generally. We also recommend best practices for the construction and use of gene catalogs in microbiome studies and highlight opportunities for future research.

In Chapter 3, we assess if the integration of long (Oxford nanopore) and short (Illumina) read sequencing of quasimetagenomes can expedite the accurate reconstruction of a bacterial pathogen genome from contaminated ice cream that caused a foodborne illness outbreak. Our results highlight that the integration of long and short read sequencing of quasimetagenomes expedites the reconstruction of a high-quality pathogen genome compared to either platform alone. Further, we show that a more complete level of information about genome structure, gene order and mobile elements can be added to the public health response by incorporating long read analyses with the standard short read WGS outbreak response. We also benchmark ten state-of-the-art short read, long read, and hybrid assembly approaches and reveal biases and areas for improvement.

In Chapter 4, we describe taxaTarget, a novel strategy for the taxonomic classification of microeukaryotes in metagenomic data. Using a database of eukaryotic marker genes and a supervised learning approach for training, we learn the discriminatory power and classification thresholds for each 20 amino acid region of each marker gene in our database. This approach provides improved sensitivity and precision compared to other state-of-the-art approaches, with rapid runtimes and low memory usage. Additionally, taxaTarget is better able to detect the presence of multiple closely related species as well as species with no representative sequences in the database. One of the greatest challenges for developing taxaTarget is the general sparsity of available sequences for microeukaryotes. We implement several algorithms, including threshold padding, to handle the missing training data and to reduce classification errors. We show how data-driven methods for learning classification thresholds from the structure of an input database can facilitate a more comprehensive analysis of metagenomic data and expand our knowledge about the diverse eukaryotes in microbial communities.

In Chapter 5, we explore the use of horizontally transferred plasmids to relate an outbreak strain to the microbiome of a suspected environmental source. Our results show how the phylogenetic analysis of individual plasmid types identifies instances recent horizontal transfer events, but how their seeming promiscuity obscures any signal about geographic location, isolation source, and time since transfer. In contrast, our statistical analysis shows that the number and combination of highly similar plasmids shared by clinical and environmental isolates supports a connection between the outbreak strain and the farms implicated by the epidemiological investigation. Together, our analysis highlights that the profile of horizontally transferred plasmids can provide information about microbial interactions, potentially connecting lineages to environments and microbiomes. However, to be useful for source tracking investigations we highlight the importance of detailed and accurate metadata, more extensive environmental sampling, and a better understanding of plasmid molecular evolution.

2. A critical assessment of gene catalogs for metagenomic analysis

This chapter was published in Bioinformatics [41] and was a joint work with Nidhi Shah (co-first author), Jay Ghurye, Alexander Stoppel, Jessica A. Goodheart, Guillermo G. Luque, Michael P. Cummings, and Mihai Pop. This project was conceived and initiated at the 2017 Bioinformatics Exchange for Student and Teachers (BEST) summer school in Heiligkreuztal, Germany. M.P. conceived this project. All authors helped initiate the project. S.C., N.S., and M.P. were involved in the design and execution of all experiments. S.C. performed most of the data analyses except for section 2.2.4 (mainly performed by N.S. and J.G.) and wrote most of the manuscript. J.G., A.S., and J.A.G. contributed to data analysis. S.C., N.S., and M.P. wrote the manuscript with contributions from all authors. All authors read, revised, and approved the manuscript.

2.1 Background

Increasingly, studies of microbial communities rely on metagenomics—the sequencing of DNA extracted directly from a microbial mixture. Assembling metagenomic reads into longer contiguous sequences (contigs) is still a computationally challenging problem, because of repeated sequences within and among genomes, uneven abundances of organisms, sequencing errors, and strain-level variation. Due to these challenges, and to limitations of sequencing technology, reconstructing complete and accurate genomes for all organisms in a single, complex metagenomic sample is still challenging. Given enough samples, metagenome assembled genomes can be reconstructed for many, but often not all, of the species comprising a microbiome. Regardless, metagenomic assemblies typically comprise many small contigs of unknown taxonomic origin.

The fragmented nature of metagenomic assemblies complicates data analysis, both because it is difficult to associate genomic fragments with individual taxa, and because it is difficult to identify related genomic fragments across samples. For these reasons, the earliest metagenomic studies focused on genes (and their inferred functions) found within assembled fragments, ignoring their precise taxonomic origin. Even in fragmented data, genes can be fairly effectively identified [42]. A gene-centric approach was used in the first large scale metagenomic study of ocean bacteria [43]. To prevent overcounting due to sequencing and assembly errors, or due to small differences in gene sequences within closely related organisms, Yooseph et al. [43] clustered the protein sequences based on similarity and focused their analysis on the representative sequence of each cluster. This gene "catalog" revealed the tremendous diversity of bacterial functions in the ocean, with the newly predicted protein sequences doubling the number of known proteins. The MetaHIT project [32] constructed a similar catalog to characterize the functional composition of the human gut microbiome. Qin et al. [30] leveraged a gene catalog as the basis for a microbiome association study in type 2 diabetes and introduced the concept of metagenomic linkage groups—groups of genes that co-vary in abundance across samples. The gene catalog thus represents the basis for grouping together genes that likely originate from a single organism, an idea further extended by Nielsen et al. [44] to help reconstruct partial genome sequences from metagenomic data.

Following these initial studies, gene catalogs have become ubiquitous in the analysis of metagenomic datasets, and have been created for the gut microbiota of multiple animals (e.g., mouse [45], rat [46], pig [47, 48], dog [49], cow [50], macaque [51], chicken [52], lion, leopard and tiger [53]), ocean bacteria [54], soil bacteria [55], and the human vagina [56] and respiratory tract [57]. Gene catalogs are commonly used to: i) reduce redundancy in the data, thereby improving estimates of diversity [29]; ii) act as a common frame of reference across samples and studies; iii) serve as a basis for metagenomic-wide association studies [58]; and iv) guide the binning of metagenomic contigs into organism-specific groups [44, 59].

Such analyses may be confounded by the specific properties of the catalog being used. Yet, to our knowledge, the structure and construction of gene catalogs have not been critically evaluated. Because the processes for constructing and using gene catalogs are broadly the same across studies, generalizable observations can be obtained from the analysis of any of the catalogs referenced above. We focus here on the Integrated Gene Catalog (IGC) [60], which seeks to provide a nearly comprehensive collection of the gene sequences identified in the human gut microbiome. We chose the IGC because it provides all the supporting metadata and intermediate files necessary to conduct a critical analysis of the structure of the resulting clusters.

2.1.1 The construction and use of gene catalogs

Catalog construction starts by identifying genes within metagenomic data. The gene sequences are then clustered together based on similarity to remove trivial differences between sequences due to fragmentary data (e.g., genes that miss the start or stop codons), sequencing errors, or small, strain-level variations. The clustering can be performed at the DNA level (e.g., the IGC [60]), or at the amino acid level (e.g., the Global Ocean Survey [43]). Analysis at the DNA level provides greater resolution for taxonomic classification, whereas the amino acid level is better suited for functional analysis and is more able to group together distantly related but functionally similar sequences. The implied, but often unstated, goal of the clustering process is to reproducibly group together sequences that have the same function and/or taxonomic origin, thereby defining the gene from which the sequences are derived in a way that is consistent across samples. Each cluster is typically represented by one sequence, either a representative selected from the sequences clustered together, or a sequence that represents the consensus of the clustered sequences. Beyond the obvious use of these sequences in a broad range of sequence-based analyses (e.g., database searches, function or structure prediction), the cluster representatives can also be used to estimate the relative abundance of the corresponding genes within microbiome samples.

2.1.2 Historical context

Clustering of biological sequences that share a common function or taxonomic origin has been at the core of biological research long before the first metagenomic experiment. Databases such as the Clusters of Orthologous Groups (COG) [61] and Pfam [62] date back to the late 1990s and were developed to organize the rapidly accumulating protein sequence information. To define the boundary of clusters, these databases used reciprocal best hit links (COG), or hidden Markov models built upon multiple alignments of related proteins (Pfam), approaches that rely on statistical significance measures instead of arbitrary thresholds based on sequence similarity. At the same time, taxonomic analyses based on housekeeping genes relied on careful phylogenetic analyses to define species boundaries [63].

In the early 2000s, metagenomic studies yielded much larger data sets than previously seen. The challenge of effectively scaling analyses to cope with increasingly larger data sets led to the development of new approaches that emphasized speed over the accuracy or comprehensiveness of the analysis. CD-HIT [31], for example, a greedy clustering approach we briefly describe below, was developed to address the challenges encountered when analyzing the data from the Global Ocean Survey. Although CD-HIT and some other clustering tools developed [64, 65] relied on fixed thresholds to determine the boundaries of clusters, it was already recognized that such thresholds were not consistent with biologically relevant entities

[66, 67]; for a given threshold, some clusters contained sequences from multiple species, whereas other species were represented in multiple clusters.

The estimation of abundances from sequencing reads is a relatively new development in metagenomic studies but has been used extensively in the study of gene expression in eukaryotes. Several factors have been identified that confound abundance estimation including multi-mapped reads, uneven depth of coverage, and sequence composition biases. Computational and statistical approaches have been developed to address such challenges [68-71].

2.1.3 Overview of the Integrated Gene Catalog

The Integrated Gene Catalog comprises 9,879,896 annotated gene clusters that were constructed from a combination of 511 prokaryotic reference genomes from species known to occur in the human gut, and 1,267 gut metagenome data sets from Chinese, American, and European cohorts. The IGC has been used to discover correlations between gut microbiome composition and resistance to immune checkpoint inhibitors in cancer patients [72], to observe that microbiome composition is modulated to a greater degree by environmental factors than by human genetics [73], to correlate glycemic response after meals with microbiome composition [74], and to identify signs of human fecal contamination in a river with sewage input [75].

The IGC was created through a multistep clustering process [60]. First, separate gene catalogs were created from the metagenomic data derived from each cohort: American (AGC), Chinese (CGC), and European (EGC), and for the sequenced prokaryotic reference genomes collection (SPGC). The three cohort-specific gene catalogs were then clustered together into a larger gene catalog called the 3CGC, which was then clustered with the SPGC catalog to create the IGC. Gene clustering was performed with CD-HIT [31]. As employed in the construction of the IGC, this tool operates in an iterative fashion, processing the gene sequences in decreasing order of length. The longest gene sequence is selected to be the representative of the first cluster. The next longest sequence is then assigned to the cluster if it matches the representative sequence with ≥ 95% sequence identity over ≥ 90% of the length of the query sequence or becomes the representative of a new cluster. In the following iterations, query sequences either become representatives of new clusters or are added to an existing cluster if they match the corresponding representative sequence sufficiently well. For most applications, only the set of representative sequences is used, however the IGC project also provides the full assignment of individual genes to clusters. Each representative gene sequence in the IGC is assigned, if possible, taxonomic and functional labels, however, only 16.3% of the sequences are assigned a genus-level annotation and only 60.4% have functional annotations.

2.2 Results

2.2.1 Inconsistent fidelity of clustering

That a 95% sequence identity cut-off is used throughout the multiple rounds of clustering in the construction of the IGC appears to imply that the final clusters are consistent with this threshold. However, the multiple rounds of clustering used to construct the IGC may yield clusters with a (much) lower identity than the intended threshold. We call this methodological artifact transitive clustering error (Figure 2.1), which occurs when different gene catalogs are sequentially clustered. Although each clustering step guarantees the 95% threshold for the sequences being clustered, this threshold does not constrain the similarity between sequences that were clustered in prior iterations. The result of transitive clustering error is an unintended increase in the effective radius of the new cluster with respect to the representative sequence (see the Appendix for a detailed explanation of transitive clustering error). When the three cohorts were clustered into the 3CGC, individual gene sequences could potentially share as low as 90% identity to the new representative gene sequences, while two sequences within a cluster may share as little as 80% sequence identity. The final clustering of the SPGC and the 3CGC, could potentially have clustered sequences with only 85% identity to the representative sequence and as low as 70% identity between sequences assigned to the same cluster.

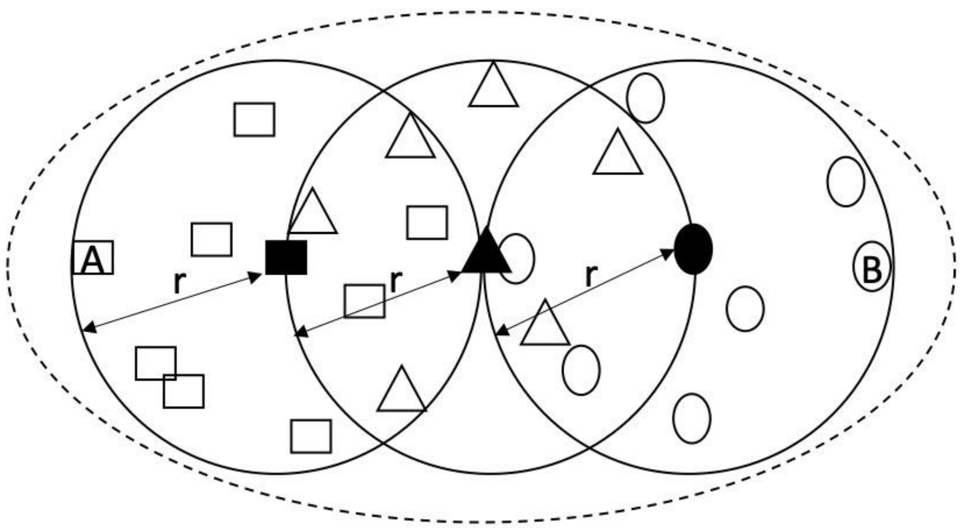

Figure 2.1 Transitive clustering error
The circles represent three clusters from three distinct catalogs. Within each catalog, the sequences within a cluster (represented by points of different shapes) are guaranteed to be within a distance r (5% divergence in the case of the IGC) from the corresponding cluster representative (solid shape). When merging multiple catalogs, only the representative sequences are clustered together (also within the same tolerance r), while the sequences contained within each cluster are implicitly assigned to the same cluster as the corresponding representative. In this figure, after clustering the representatives in one round, the triangle cluster representative is the representative of a meta-cluster (dashed line) that includes the representative sequences of the square and circle clusters. Within this cluster, the maximum distance between two sequences (marked with A and B in the figure), may be as high as 4r, or 20% sequence divergence in the case of the parameters used in the IGC. The distance between a sequence and its corresponding cluster representative may be as high as 2r, or 10% sequence divergence.

To evaluate the actual impact of transitive clustering error within the IGC, we focused on the 255,191 IGC gene clusters that contained at least 100 sequences each. Among these clusters, 29.6% contained sequences that differed from the cluster representative by more than the intended 95% identity cut-off (Figure 2.2A).

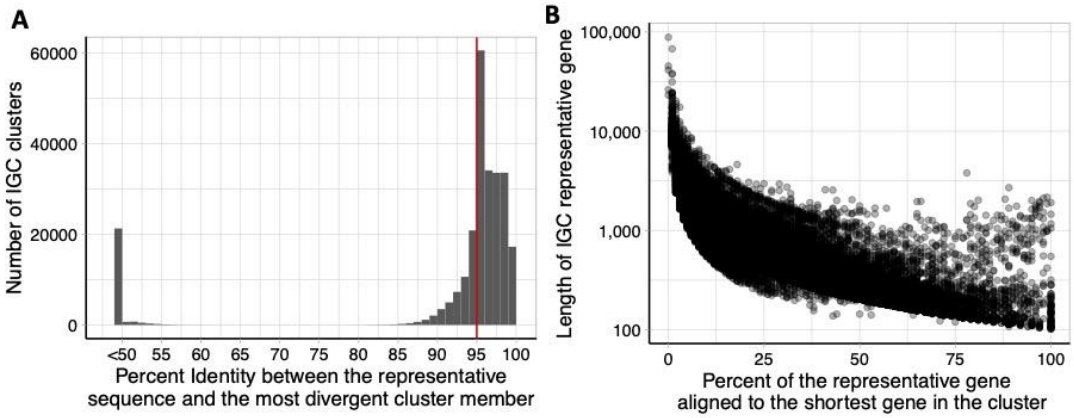

Figure 2.2 Undesirable effects of gene clustering in the IGC
Data shown refer to the 255,191 IGC clusters that contain 100 sequences or more. A) The distribution of CD-HIT percent identity between the representative and the most divergent cluster member. The vertical red line indicates the 95% identity clustering threshold used to create the IGC. Note that many sequences are below the target threshold of 95%. B) Relationship of percent of the representative gene aligned to the shortest cluster member and the length of the representative gene.

Furthermore, 8.2% of the clusters contained sequences that are different by 50% or more from the corresponding cluster representative. This difference is much higher than the expected error due to transitive clustering. An explanation is that the construction of the IGC did not require full length alignments to each cluster representative, but rather allowed matches that cover as little as 90% of the clustered sequence. In the worst case, after two or more rounds of clustering, sequences within an IGC cluster may not overlap with the selected representative sequence at all (Figure 2.3).

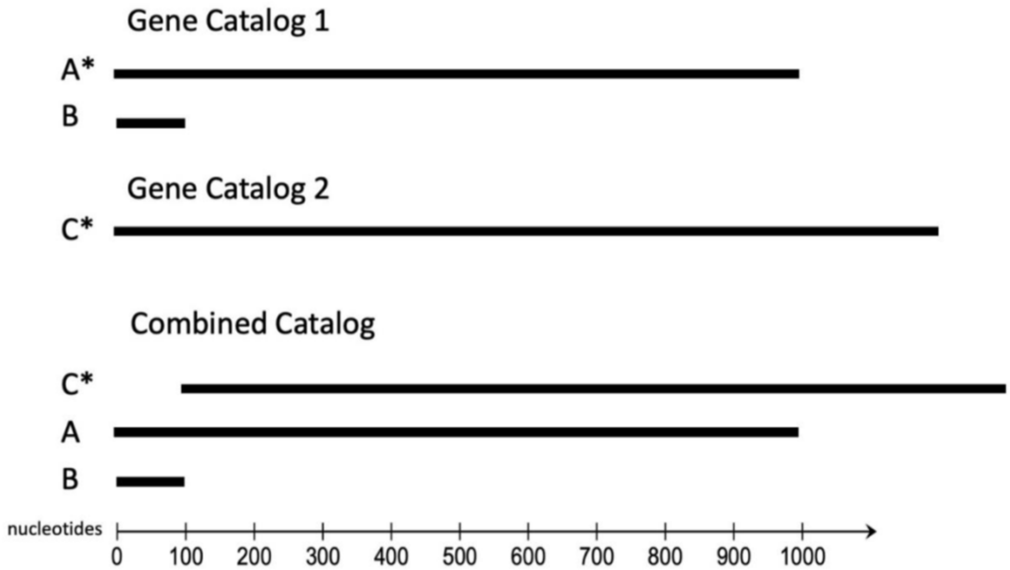

Figure 2.3 Clustering separate gene catalogs with CD-HIT can recruit sequences that do not overlap with the representative sequence

A schematic example of how, in a worst-case scenario, clustering separate gene catalogs with CD-HIT can recruit sequences that do not overlap with the representative sequence given the IGC clustering parameters. The sequences within each gene catalog are aligned. Here * denotes the representative sequence of the catalog. Gene A and Gene B were clustered together to create Gene Catalog 1. Gene A is the representative sequence because it is the longest sequence (default of CD-HIT). In this case 100% of the length of Gene B aligns to 10% the length of Gene A with 100% identity. Gene C is a representative sequence in Catalog 2 with no clustered sequences. Gene A and Gene C were clustered to create the Combined Catalog. Gene C becomes the new representative, because it is longer than Gene A, and Gene A and Gene B become cluster members. In the Combined Catalog, 90% of the length of Gene A aligns to Gene C with 100% identity and Gene B has no overlap with Gene C at all.

The process used to construct the IGC does not constrain the fraction of the representative sequence that needs to match the sequences within the cluster. This choice makes it possible for two sequences to both align to the cluster representative perfectly without sharing any sequence with each other. As an example, cluster 303 contains four sequences of different lengths—16,111 nt (representative), 7,122 nt, 3,012 nt, and 2,982 nt. All these genes are complete, spanning from start codon to stop codon, and originate from the SPGC (genes found in nearly complete reference genomes). The alignment between the three genes to the cluster representative (Figure 2.4) demonstrates the lack of overlap between the individual sequences, suggesting that they align to distinct domains of the representative sequence, rather than representing variants of this gene. This artifact may be wide-spread within the IGC—within the 255,191 IGC clusters with a minimum of 100 members, the mean difference between longest and shortest gene length is 590 nt, representing an average of 14.4% of the length of the cluster representative (Figure 2.2B).

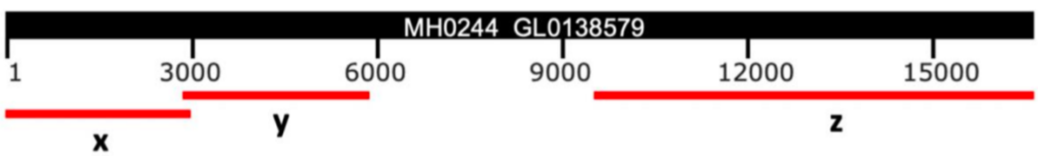

Figure 2.4 BLASTN alignment of IGC Cluster 303

BLASTN alignment of the IGC Cluster 303 representative sequence, MH0244_GL0138579 (16,611 nt), and the 3 cluster members x (469585.HMPREF9007_02027, 2,982 nt), y (469585.HMPREF9007_02028, 3,012 nt), and z (469585.HMPREF9007_02029, 7,122 nt). All were predicted as complete genes (from start to stop codon), yet each cluster member only partially aligns to the representative with a small overlap between x and y and no overlap between y and z.

2.2.2 Taxonomic inconsistency of clusters

The 95% identity threshold selected by the IGC was intended to create clusters with taxonomic homogeneity at the species level [60]. Taxonomic homogeneity is desirable for analyses with the IGC, however, as we briefly described above, it has long been recognized that no specific threshold can universally and accurately capture biologically meaningful boundaries [66, 67].

This can be demonstrated by clustering the genes of genera like *Bacteroides* and *Lactobacillus* comprising multiple species within which many strains have been sequenced. We separately clustered the RefSeq genes from 167 *Bacteroides* (5,355,696 genes) and 166 *Lactobacillus* species (1,876,284 genes) using the IGC clustering parameters. Of the resulting

438,106 *Bacteroides* and 256,949 *Lactobacillus* clusters, 32% and 24% were composed of multiple species, respectively.

The taxonomic homogeneity of the IGC clusters can be most readily assessed within the SPGC because this gene catalog has well-defined taxonomic labels; however, we note that the SPGC only contains 200 species with sparse representation per species (a mean of 2.6 reference genomes). Still, we found that 42,208 (6.4%) of all clusters in the SPGC grouped together sequences from multiple distinct species, with a maximum of 21 species in a single cluster.

To estimate the number of species within the IGC clusters derived from sequences with unknown taxonomic origin (namely, the three country-specific catalogs), we focused on a subset of 200 IGC clusters: the 100 largest clusters and 100 randomly chosen clusters from those with at least 100 sequences each. We aligned each sequence within an IGC cluster to the NCBI nr database (version 5) using Diamond [76] (version 0.9.29). We used the same alignment thresholds as those used by the IGC, requiring at least 95% sequence identity and 90% query coverage. We retained all database entries that matched each query sequence within these thresholds. We conservatively inferred the number of species per gene cluster using a minimum set cover approach. Specifically, we identified the smallest number of species such that each sequence had at least one hit to a database sequence from one of these species. As seen in Figure 2.5A, 73% of clusters (57% of the largest and 89% of the randomly selected clusters) are covered by a single species. If we used just the top database hit for each sequence, the most commonly used approach in practice, only 20.5% of clusters (5% of the largest and 36% of the randomly selected clusters) were composed of a single species (Figure 2.5B).

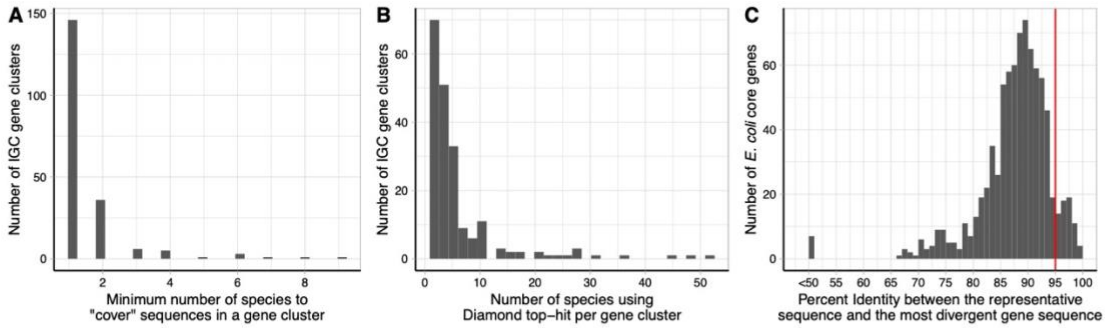

Figure 2.5 IGC clusters can contain more than one species

A and B show the taxonomic heterogeneity of 200 IGC clusters (the 100 largest clusters and 100 randomly chosen clusters with at least 100 sequences). A) The minimum number of species such that each sequence in a cluster had at least one significant Diamond hit to one of these species B) Number of species per cluster if each sequence is assigned the label of the top Diamond result. C) The distribution of CD-HIT percent identity between the representative and the most divergent gene sequence for the 818 core genes of *Escherichia coli* identified from 86,830 assemblies. The red vertical line denotes 95% identity, the IGC clustering threshold.

To explore the converse, the possibility that variants of a gene from a single species may be distributed across multiple clusters, we analyzed a collection of 86,830 *Escherichia coli* genomes obtained from the GenomeTrakr database [38]. When focusing on just the 818 core genes of the *E. coli* pan-genome (genes found in all the genomes), the mean sequence identity between the representative and the most divergent clustered sequence was 87.7% which is lower than the 95% threshold used by the IGC. In fact, only 63 core genes met or exceeded the 95% threshold and would have been clustered properly by the IGC (Figure 2.5C).

2.2.3 Hidden species within the IGC

A direct consequence of multi-species clusters is the possibility that genes from an individual species may be “hidden” by representative sequences belonging to a different species. A species for which no gene is selected as a representative for a cluster in the catalog becomes effectively undetectable in the samples being analyzed.

To explore the extent of this problem, we focused on just the SPGC (genes from complete and near complete genomes) because these genes have well defined taxonomic labels. Within the SPGC, the number of representative genes per species ranged from 139 (*Escherichia sp. 1_1_43*) to 28,404 (*Escherichia coli*). We simulated reads from 507 genomes from the same species (or strain, if known) as the SPGC reference genomes, and mapped these reads to the SPGC using Bowtie2 [77]. As expected, the rate of assigning reads to a species was correlated with the number of representative genes for the species (Figure 2.6). A possible confounding factor might be the fraction of reads that map ambiguously to multiple species, however the median fraction of multi-mapped reads was only 3% across species. Only 129 of the 201 species in the SPGC had an assignment rate of 90% or higher, i.e., 90% of the reads originating from these genomes would be assigned a correct species-level taxonomic label. At one extreme, *Escherichia sp. 1_1_43*, had the lowest number of representative genes and the lowest assignment rate at 2%. Despite having many representative genes, *E. coli* only had an assignment rate of ~83%, because of the large number of closely related species in the SPGC. All four *Shigella sp.* within the SPGC had low assignment rates: 17%, 11%, 8% and 7% for *S. flexneri*, *S. dysenteriae*, *S. boydii*, *S. sonnei*, respectively. This is because the reads from *Shigella sp.* often map to clusters with an *E. coli* representative sequence.

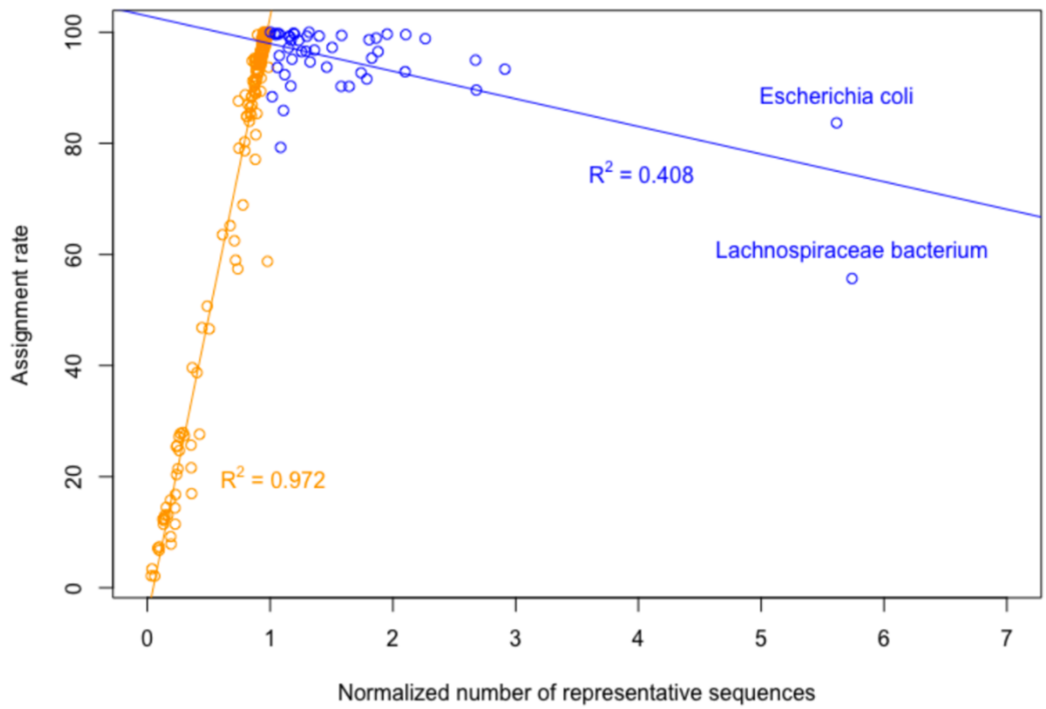

Figure 2.6 The relationship between the number of representative genes per species and the assignment of simulated reads

The relationship between the number of representative genes (normalized by the mean number of genes per genome) per species and their assignment rate in a simulated metagenomic dataset of the SPGC genomes. The assignment rate is the percent of simulated reads from a species that map to the corresponding representative sequences for that species in the SPGC. For most species in the SPGC, the number of representative genes (normalized by the mean number of genes per genome) is 1 or less (orange). The assignment rate for these species has a positive correlation (orange least squares line) with the number of representatives. For some species, however, the number of representative genes normalized by the mean number of genes per genome can be greater than 1 (blue). These species have genes from multiple genomes and are effectively represented as a pangenome in the SPGC. For example, *E. coli* has 28,404 representative genes and 124 genomes in the SPGC. For these species there is a weak negative correlation (blue least squares line) between the assignment rate and the number of representatives.

Due to the importance of *Shigella sp.* for human health, we further analyzed 20 known virulence/toxin genes of *S. sonnei* [78-80] (Table 2.1). Only 11 of the 20 genes were taxonomically labelled as *Shigella*, seven were labelled as *Escherichia*, and two, set1A and set1B, were not found at all. Notably, Shiga toxins Stx1A and Stx1B are labelled as *Escherichia*, even though they are part of a mobile prophage genome, which has been horizontally transferred among many *Enterobacteriaceae* [81], highlighting the difficulty of annotating a mobilome.

Table 2.1 Taxonomic annotation of twenty virulence/toxin genes of *Shigella sonnei* when aligned to the SPGC catalog.

Toxin/Virulence factor	Genus of most similar gene in IGC	Percent identity of the top BLAST hit in IGC
ShiA	*Shigella*	100.00
ShiB	*Shigella*	94.41
ShiC	*Shigella*	100.00
ShiD	*Shigella*	100.00
ShiE	*Shigella*	99.43
ShiF	*Shigella*	99.75
ShiG	*Escherichia*	84.44
IucA	*Escherichia*	99.83
IucB	*Escherichia*	99.37
IucC	*Escherichia*	96.38
IucD	*Shigella*	99.78
IutA	*Escherichia*	99.45
Pic	*Shigella*	99.64
GtrA	*Shigella*	99.34
GtrB	*Shigella*	98.15
SigA	*Shigella*	97.67
set1A	Not found	NA
set1B	Not found	NA
Stx1A	*Escherichia*	100.00
Stx1B	*Escherichia*	100.00

2.2.4 Using the IGC as a reference for metagenomic analyses—simulated data

The primary strategy for using the IGC as a reference when analyzing metagenomic data sets involves mapping sequencing reads to the representative sequences of the clusters. Although a seemingly straightforward bioinformatics task, the selection of mapping tools, parameters of the mapping process, and characteristics of the reads themselves (e.g., read length) may have a significant impact on the results. To evaluate the effects of such features on the use of the IGC for metagenomic analysis, we simulated three metagenomic samples composed of the species in the SPGC. Two samples simulated Illumina reads (100 nt, 250 nt), and the other simulated 454/IonTorrent reads (225 nt). We compared mapping statistics for tools that are widely used in metagenomic analyses, BWA-MEM [82] and Bowtie2 [83] with default parameters, and BLASTN [84] with thresholds of 95% identity, 90% read coverage, and default values for all other parameters (Table 2.2).

Table 2.2 Mapping statistics when aligning simulated reads to the SPGC

Read datasets	BLASTN	Bowtie2	BWA-MEM
Illumina 100 nt	74.31	86.44	96.22
Illumina 250 nt	43.98	76.49	98.97
454 Roche 225 nt (mean)	64.48	77.82	98.18

The fraction of reads mapped by different tools, and across different read lengths, varied substantially (Table 2.3). BLASTN consistently mapped fewer reads than the other tools. The gene abundance profiles estimated from these mappings differed significantly across different mapping tools (Mann Whitney U test, p-value < 0.001) at every read length, suggesting the choice of mapping tool may confound abundance estimates and, therefore, the associations derived from the data (Table 2.4). Furthermore, nearly half of the reads multi-mapped, i.e., mapped equally well to multiple IGC clusters. Multi-mapped reads can confound taxonomic classification and estimates of abundance, as previously highlighted in RNA-seq studies [85].

Our results suggest the need for abundance estimation algorithms that can account for mapping ambiguity [68-70], which are rarely used in metagenomic studies.

Table 2.3 Read mapping when using different alignment tools

Mapping tool	Dataset	Unmapped reads	Reads mapped exactly once	Multi-mapped reads	Total reads
BLASTN	454 Roche 225 nt	12046662 (35.52%)	20607264 (60.76%)	1259937 (3.72%)	33913863
Bowtie2	454 Roche 225 nt	7523531 (22.18%)	12727602 (37.53%)	13662730 (40.29%)	33913863
BWA-MEM	454 Roche 225 nt	615080 (1.81%)	17789182 (52.45%)	15509601 (45.73%)	33913863
BLASTN	Illumina 100 nt	24590586 (25.69%)	63782504 (66.64%)	7339930 (7.67%)	95713020
Bowtie2	Illumina 100 nt	12977730 (13.56%)	42142225 (44.03%)	40593065 (42.41%)	95713020
BWA-MEM	Illumina 100 nt	3618165 (3.78%)	49637777 (51.86%)	42457078 (44.36%)	95713020
BLASTN	Illumina 250 nt	21407600 (56.03%)	16112369 (42.17%)	690019 (1.81%)	38209988
Bowtie2	Illumina 250 nt	8984244 (23.51%)	14631373 (38.29%)	14594371 (38.20%)	38209988
BWA-MEM	Illumina 250 nt	392069 (1.03%)	20811950 (54.47%)	17005969 (44.50%)	38209988

Table 2.4 P-values from Mann Whitney U Test comparing the gene abundance profiles.

Read Length	BWA-MEM vs Bowtie2	BWA-MEM vs BLASTN	Bowtie2 vs BLASTN
Illumina 100 nt	2.68×10^{-251}	1.12×10^{-39}	2.45×10^{-25}
Illumina 250 nt	3.27×10^{-07}	0.0	3.84×10^{-123}

Together, multi-mapped reads and the poor visibility of some species within the catalog, led to ~20% of the reads mapping to gene clusters classified as a different genus than that from which the reads originated (Table 2.5). This raises concerns about the accuracy of taxonomic profiles derived from real metagenomic data given that these reads were generated from the genomes used in the construction of the IGC.

Table 2.5 Read mapping statistics for testing the taxonomic classification performance of the IGC on data simulated from the SPGC.

Read Length	Percent of reads mapped to IGC	Percent of reads mapped to correct genus
Illumina 100 nt	86.4	81.7
Illumina 250 nt	76.5	82.1

2.2.5 Using the IGC as a reference for metagenomic analyses—real data

In addition to the read mapping artifacts discussed previously, genes that are not represented in the IGC but are present in a sample can confound the analysis of metagenomic data. Prior studies have demonstrated the IGC is not a comprehensive representation of the diversity of the human gut microbiome, lacking many genes found in the gut of infants [86], patients suffering from various diseases such as gout [87] or diabetes [88], adults from India (only 61% of their gene catalog mapped to the IGC) [89], and even adult twins from the UK (in which a putative 1.5 million genes were not present in the IGC) [90].

To investigate how read mapping artifacts and genes not represented in the catalog impact analyses based on the IGC, we used a human gut sample from a 61 year-old Cameroonian male with a hunter gatherer diet (SRA accession ERR2619707) [91]. We assembled the data with MEGAHIT [92] and predicted genes using Prokka [93]. Only 66.6% of the predicted genes from this sample clustered to an IGC gene representative, genes to which we refer as the clustered predicted genes. The other genes predicted from the sample could not be confidently assigned to IGC clusters (and thus are likely not represented in the IGC), and we refer to these genes as the unclustered predicted genes.

We separately mapped the reads from the Cameroon data set with Bowtie2 to the two sets of genes predicted from the sample and the IGC clusters, respectively (Figure 2.7). The percent of reads mapping to the predicted genes and the IGC was similar (59.0% to the predicted genes and 55.3% to the IGC), but the percent of multi-mapped reads was much higher for the IGC (24.1%) compared to the predicted genes (3.8%). The reads also mapped to an order of magnitude more IGC clusters (1,369,981) than predicted genes (177,745). Together this suggests a high false positive rate, i.e., that reads from unclustered predicted genes are mapping to IGC clusters representing potentially unrelated genomic sequences and/or functions.

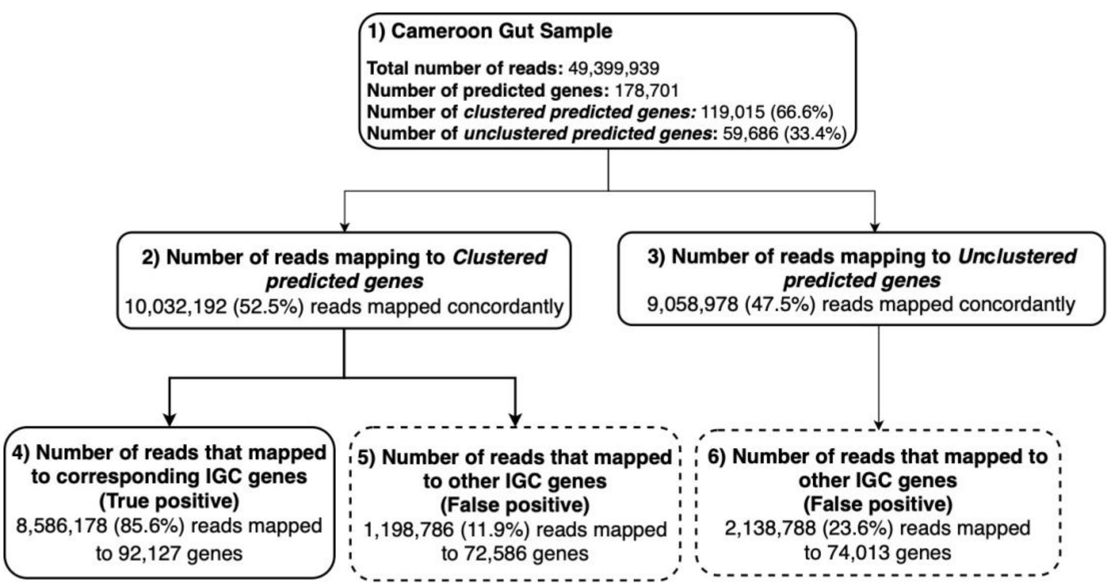

Figure 2.7 Analysis of reads from Cameroonian human gut metagenome

Analysis of reads from Cameroonian human gut metagenome sample. Box 1 shows the general statistics of the sample. 66.6% of the predicted genes could be assigned to IGC gene clusters clustered predicted genes. 33.4% of the predicted genes could not be confidently mapped to the IGC clusters unclustered predicted genes. 19,091,170 reads mapped concordantly to the predicted genes (52.5% to the clustered predicted genes and 47.5% to the unclustered predicted genes). Among the 10,032,192 reads that mapped concordantly to the clustered predicted genes (Box 2), 11.9% mapped to a different IGC gene than expected (false positives denoted by dashed line Box 5). Of the 9,058,978 reads that mapped concordantly to the unclustered predicted genes (Box 3), 23.6% mapped to IGC genes (false positives denoted by dashed line Box 6).

To determine the IGC clusters to which the reads from the clustered predicted genes and the unclustered predicted genes were aligned, we focused our analysis on the read pairs that mapped concordantly to both the predicted genes and to the IGC clusters (24.1% of all reads). A read pair is considered concordantly mapped when the forward and reverse reads of the pair map to a gene with the correct insert size and orientation. Such concordant mappings are less likely to represent mapping artifacts. Given that each clustered predicted gene has a corresponding IGC cluster, we would expect the reads mappings to also be shared between the gene and the cluster to which it is related. Among the 10,032,192 reads that concordantly mapped to clustered predicted genes, 11.9% mapped to a different IGC gene than expected. Conversely, we would expect few read pairs which map to the unclustered predicted genes to map to any IGC clusters given that these genes do not share sufficient similarity with any IGC cluster. Of the 9,058,978 reads that concordantly mapped to the unclustered predicted genes, 23.6% mapped to IGC genes (Figure 2.7).

2.2.6 Analysis of other gene catalogs

A survey of 24 gene catalog studies from the last few years highlights that many were created using a similar clustering algorithm as the IGC and thus likely share many of the same issues as those identified above (Table 2.6). While none of these catalogs provided all the necessary metadata and intermediate files to perform the same analyses as done for the IGC, we were able to predict which issues likely affect the catalogs based upon the description of the methods used to construct these catalogs. We note that 5 of the 24 catalogs were affected by transitive clustering error. Additionally, at least 15 catalogs allowed genes of highly divergent lengths to be clustered together. Further, taxonomic inconsistency and hidden species also likely affect 23 of the catalogs.

Table 2.6 Non-exhaustive list of microbial gene catalogs and the issues affecting them.

Gene catalog	Year published	Transitive clustering error	Clusters sequences of highly different lengths	Taxonomic inconsistency	Hidden species	Clustering criteria
Human gut (cirrhosis) [94]	2014	Yes	Yes	Yes	Yes	- Pairwise comparison of all genes with BLAT: >95% identity and >90% of the shorter gene length. - Merged genes from three catalogs using the same clustering technique.
Mouse gut [45]	2015	No	Unclear	Yes	Yes	Pairwise comparison of all genes with BLAT: >95% identity and overlap >90%
Human gut (infants) [86]	2015	No	Yes	Yes	Yes	Pairwise comparison of all genes with BLAT: >95% identity and >90% of the shorter gene length
Human gut (diabetes) [88]	2015	No	Yes	Yes	Yes	Predicted protein-coding genes with a minimum length of 100 bp were clustered at 95% sequence identity using CD-HIT with parameters set to: -c 0.95, -G 0 -aS 0.9, -g 1, -r 1.
Pig gut [48]	2016	No	Unclear	Yes	Yes	Pairwise comparison of all genes with BLAT: >95% identity and overlap >90%
Human gut (gout) [87]	2016	No	Yes	Yes	Yes	Genes were clustered with CD-HIT using a sequence identity cut-off of 0.95 and a minimum coverage cut-off of 0.9 for the shorter sequences.
Human gut (diabetes) [90]	2016	Yes	Yes	Yes	Yes	- Genes were clustered using CD-HIT of the MOCAT pipeline (95% identity, 90% overlap) - Merged the gene set with the IGC catalog using CD-HIT
Chicken gut [52]	2018	Unclear	Yes	Yes	Yes	Gene catalog was constructed using CD-HIT-EST with parameters set to: -c 0.95 -n 10 -G 0 -aS 0.9
Rat gut [46]	2018	No	Yes	Yes	Yes	Gene ORFs were clustered using CD-HIT with a criterion of 95% identity >90% of the shorter ORF length with default parameter except “-G 0 –n 8 –aS 0.9 –c 0.95 –d 0 –g 1”
Dog gut [49]	2018	No	Unclear	Yes	Yes	Genes were clustered at 95% identity using CD-HIT
Macaque gut [51]	2018	No	Unclear	Yes	Yes	Pairwise comparison of all genes using CD-HIT with identity of >95% and overlap of >90%
Human gut (children) [95]	2018	Yes	Yes	Yes	Yes	- Clustered gene based on sequence similarity at 95% identity and 90% coverage of the shorter sequence using CD-HIT - Merged with the IGC using the same CD-HIT clustering technique to form a comprehensive catalog
South China soil [55]	2019	No	Unclear	Yes	Yes	Nucleic acids longer than 100bp were translated into amino acid sequences. Pairwise comparison of all genes using CD-HIT with parameters >95% identity and >90% overlap
Pig gut [96]	2019	Yes	Unclear	Yes	Yes	- Predicted genes were clustered at the nucleotide level using CD-HIT with >95% identity and >90% overlap - Combined the catalog with an earlier Pig gut catalog to create a comprehensive catalog
Human lung [57]	2019	No	Yes	Yes	Yes	Genes with a length ≥100 bp and without Ns (unidentified nucleotides) were selected to construct non-redundant gene sets using CD-HIT with criteria of >95% identity and >90% alignment of shorter sequence (–c 0.95 –aS 0.9)
Human gut (Indian cohort) [89]	2019	Yes	Unclear	Yes	Yes	- Pair-wise alignment of genes using BLAT and the genes that had an identity >95% and alignment coverage >90% were clustered into a single set of non-redundant gene. - The gene catalog constructed from Indian samples was combined with the IGC to construct a non-redundant gene catalog (using identity ≥95% and alignment coverage ≥90%)
Rat gut [97]	2019	No	Yes	Yes	Yes	Predicted ORFs were clustered using CD-HIT with criteria of >95% identity and >90% alignment of shorter ORF. (-c 0.95, -G 0, -aS 0.9, -g 1, -d 0)
Panthera gut [53]	2020	No	Yes	Yes	Yes	Predicted genes from contigs and from the top abundant microbial species were clustered using CD-HIT using a sequence identity cutoff of 0.95 and minimum coverage cutoff of 0.9 for shorter sequences
Cow gut [50]	2020	No	Yes	Yes	Yes	Predicted genes were clustered using CD-HIT with ≥95% identity and ≥90% overlap of the shorter sequence (-n 8 -d 0 -g 1 -T 6 -G 0 -aS 0.9 -c 0.95)
Mouse gut [98]	2020	No	No	Unclear	No	Predicted ORFs were taxonomically annotated at different levels, i.e., ORF, contig, and bin
Human dental caries [99]	2020	No	No	Yes	Yes	Predicted ORFs were clustered using CD-HIT default parameters
Human vagina [56]	2020	No	Yes	Yes	Yes	Genes and gene fragments that were at least 99 bp long, with greater than 95% identity over 90% of the shorter gene length were clustered together by CD-HIT-EST
Rhizosphere soil [100]	2020	No	Yes	Yes	Yes	Predicted genes were clustered using CD-HIT with >95% sequence identity
Sheep rumen [101]	2020	No	Yes	Yes	Yes	CD-HIT tool with the similarity threshold of 95% was used to remove redundant genes

2.3 Discussion

Gene catalogs help organize the vast volumes of data generated in metagenomic experiments. If carefully constructed, they provide a valuable resource for the analysis of metagenomic samples. Through our analysis of the IGC — one of the largest gene catalogs available to scientists today — we have highlighted how the design and construction of a gene catalog can affect downstream analyses in unintended ways. These issues affected a large percent of the gene catalogs we found in the literature because many were constructed using similar methods as the IGC.

Perhaps the most prevalent and important source of error for gene catalogs is caused by clustering gene sequences with a fixed threshold, creating clusters composed of sequences with variable levels of taxonomic relatedness. Our observation recapitulates the finding that no specific sequence similarity threshold can be used to consistently capture a particular taxonomic level or functional category. This finding has been well documented previously in the context of 16S rRNA sequencing [66, 67]. Clustering in this manner effectively hides the taxonomic origin of all but the gene sequences selected as cluster representatives. As a result, each species in a catalog might have a different proportion of genes that are not represented (that are hidden by the genes of other species), genes that are represented once and genes that are represented in multiple copies. This can introduce bias in downstream analyses that aim to explore the presence or abundance of taxa across samples, a bias already noted in the community [102]. For example, if a catalog contains multiple variants of a gene from a species, metagenomic reads from that gene and species might map to multiple variants in the catalog either uniquely or by multimapping. Through our analysis of the hidden species of the SPGC and the E. coli core genes, we have shown that this effect is non-uniform across taxonomic groups and can result in the biased recruitment of reads across taxa.

Another common source of error for gene catalog construction is the clustering of genes of widely different lengths. This can result in clusters where there is little or no overlap between cluster members. While it is not currently possible to confirm the functional consistency of all clusters in a gene catalog, if cluster members share little sequence similarity with the representative (which is treated as the functional homolog of all cluster members) it is likely that they do not share the same function. Furthermore, assessing the relationship between sequence and functional similarity is non-trivial [103] even in the absence of the confounding information introduced by the co-clustering of sequences with widely-divergent lengths.

The iterative clustering of catalogs can further exacerbate all the previously mentioned issues by amplifying the differences between sequences assigned to a cluster. Among the gene catalogs we have explored (Table 2), the use of a multi-step clustering process is typically used for two purposes: to mitigate computational costs, and/or to update an old catalog by merging it with a newer one. However, none of the studies we analyzed considered the amount of error introduced by iterative clustering. It is certainly desirable to develop computationally efficient catalog construction methods as data sets increase in size, as well as to efficiently incorporate new data into existing catalogs. Our analysis, however, suggests that it is important to ensure that the fidelity of the clusters is not impacted by computational convenience, and highlights the need for additional research in this field.

Coupled with the issues arising from the structure of the clusters themselves, we have shown that the use of the IGC to analyze a real metagenomic sample induces many analytical artifacts, including a high false positive rate — IGC clusters that are not actually found in a sample, but which “recruit” many reads nonetheless. Conversely, as the number of species and the number of their gene variants represented in the catalog increases, so will the number of reads that map ambiguously [104]. As a result, using gene catalogs that are constructed similarly to the IGC for metagenomic studies will likely introduce analytical artifacts that outweigh the benefit of the common frame of reference these catalogs provide.

While raising these concerns, we agree with the authors of the IGC that properly constructed gene catalogs can be an effective reference for metagenomic studies. However, to maximize their usefulness, gene catalogs should either be created directly from the samples being analyzed or from closely related samples. Our findings indicate that the goal of tracking individual clusters across studies is not met by the IGC and other similarly constructed catalogs. We believe that universal taxonomic identifiers and gene ontologies represent a better approach for relating findings across gene catalogs and metagenomic studies. For gene catalogs to be used as global resources for metagenomic data analysis, new methods for updating catalogs and accounting for biases introduced by read mapping tools needs to be researched. For now, we believe the best use case for gene catalogs is within the narrow context of the samples used to create them.

Our results highlight pitfalls that need to be avoided when constructing such catalogs and reveal several best practices:

(1) The iterative integration of clusters should be avoided as it amplifies the errors inherent to the clustering process. A multi-step clustering process may be necessary to mitigate computational costs, however we recommend limiting the number of rounds and accounting for the growth in cluster diameter that is due to the multi-round process.

(2) Arbitrary similarity thresholds should be avoided, and instead researchers should use approaches that are able to dynamically tune clustering parameters [67, 105-108].

(3) The clustering procedure should ensure all sequences within a cluster are of similar length.

(4) The construction of gene catalogs should not exclusively rely on data from metagenomic experiments, but rather should be augmented with genomic sequences from organisms that are commonly found at low abundance in the samples of interest (including eukaryotes and viruses), as such organisms are unlikely to be assembled sufficiently well within the metagenomic data.

(5) The alignment of sequences to the catalog, as well as estimation of gene abundances from the alignments, should be conducted in a way that adequately addresses non-specific mapping. Several approaches have been developed for RNA-seq analysis that effectively handle multi-mappings in an alignment-free manner [68-70], though it remains to be seen whether these are sufficiently effective in metagenomic settings or whether the underlying algorithms need to be adapted.

During the preparation of our manuscript, a new catalog was published [109], which partly addresses some of the issues we have highlighted above. The underlying data being clustered were derived from cultured genome sequences and metagenome-assembled sequences, potentially ensuring a higher quality protein catalog (the Unified Human Gastrointestinal Protein catalog). Gene-level clustering was performed at the protein level in one round of clustering, thereby avoiding transitive clustering error. Notably, the authors of this new study re-clustered the genes from the IGC and appear to be unaware of the blow-up in divergence caused by the iterative process used by the IGC: “We clustered the IGC only at 90% and 50% protein identity, as it was originally de-replicated at 95% nucleotide identity” [109]. The Unified Human Gastrointestinal Protein catalog was provided as multiple catalogs constructed with different similarity thresholds, acknowledging that no threshold is appropriate for all analyses. Some of the pitfalls identified above, however, still apply to the new catalog. When clustering protein sequences, Almeida et al. only control the fraction of the clustered sequence that needs to match the cluster representative (80% in this case), raising the possibility of artifacts such as that highlighted in Figure 2.4. Furthermore, the new catalog includes the Unified Human Gastrointestinal Genome catalog which is constructed in a two-step process to address the computational cost of clustering. The paper does not indicate that the authors are aware of the additional sequence divergence introduced by this process.

A full-fledged analysis of the new catalog, similar to what we have described above, is beyond the scope of this manuscript. However, as discussed here, it is apparent that issues such as those we have described are not widely appreciated in our community. We hope that our manuscript provides readers with an appreciation for the complexity of sequence clustering, particularly as it relates to metagenomic sequence analysis, and leads to a more thoughtful consideration of the pitfalls we have identified when using gene catalogs as a reference for data analysis.

2.4 Appendix

2.4.1 Transitive clustering error

Clustering a large number of sequences can require impractical amounts of computing time and memory. One technique for addressing the computational cost of clustering uses a divide and conquer paradigm: disjoint subsets of the sequences are clustered separately, then the cluster representatives are clustered together in one or more additional rounds of clustering. When representative sequences from different subsets are clustered together, all members of the corresponding clusters implicitly become part of the resulting cluster. In the IGC, this process was conducted in three rounds. Sequences from each of three distinct geographic regions (American, Chinese, and European) were clustered separately, as were sequences extracted from isolate genomes within the NCBI and EMBL databases, resulting in four distinct catalogs: AGC, CGC, EGC, and SPGC, respectively. In a second round, the three geographically defined catalogs were clustered together, yielding a new catalog, 3CGC, which were then clustered together with the SPGC in a third round of clustering. Each round of clustering used the same cut-offs for the percent identity between a sequence and the cluster representatives, and for the fraction of the sequence that needs to align to the cluster representative for it to be assigned to a cluster.

In the following, we discuss the implications of using such an iterative clustering process on the size of the resulting clusters. We focus on two measures of the "tightness" of clusters: the radius (maximum distance between a sequence and the cluster representative); and the diameter (maximum distance between two sequences within a cluster). For the purpose of this discussion, we ignore the impact of partial alignments between a sequence and the cluster representative and focus exclusively on percent identity as a measure of distance between sequences.

The percent identity cut-off provided to CD-HIT controls the radius of the clusters. After a single round of clustering, the maximum effective radius, R, of the clusters is the same as the cut-off, r, that was given as a parameter to CD-HIT. The maximum effective diameter, D, is exactly 2r. Below, we will show that, with each round of clustering, both the maximum effective radius and diameter of the resulting clusters increases despite using the same cut-off, r, when clustering the representative sequences of clusters generated in a prior round. We call this unintended increase in the effective radius and diameter of clusters transitive clustering error.

2.4.2 A general formulation for transitive clustering error

We will start by assuming a set of clusters already constructed, C_1, C_2, ..., C_n, which have the effective radii R_1, R_2, ..., R_n. We explore here the impact of clustering together the representative sequences of the clusters contained in C_1, C_2, ..., C_n, as defined by the effective radius and diameter of the resulting clusters. Before we proceed, it is important to note that our analysis focuses on the worst-case scenario, i.e., we show that it is possible that at least one of the resulting clusterings can have the values for R and D as defined below (Figure 2.8). Whether such a worst-case situation may occur depends on the characteristics of the data.

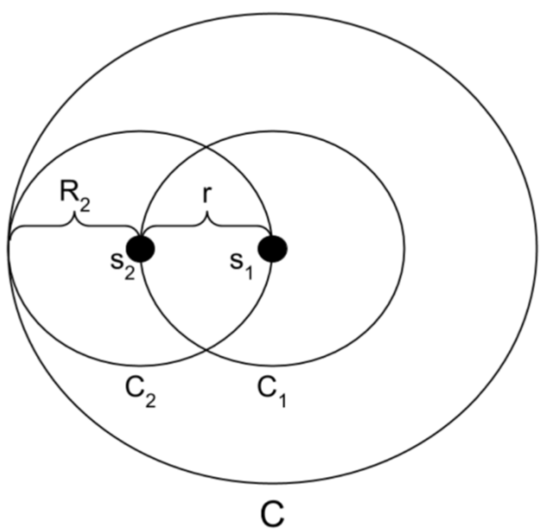

Figure 2.8 Maximum effective radius when clustering two representative sequences
An example cluster, C, with the maximum possible effective radius when clustering the representatives from C1, C2, ..., Cn with tolerance r.

Lemma 1: The maximum effective radius, R, of the resulting clusters is,

$R = \max(R_1, R_2, ..., R_n) + r$

Proof: Refer to Figure 2.8. Without loss of generality, we can assume one resulting cluster, C. By definition, the representative sequence of this cluster must be the representative sequence of one of the clusters in $C_1, C_2, ..., C_n$. Without loss of generality, we assume that this is the same as the representative sequence s_1 of the cluster C_1. Further, assume that there is another cluster whose representative sequence was clustered together with s_1. This cluster cannot come from the same catalog as C_1 since its representative sequence is within distance r of s_1 and thus would have been clustered with s_1 already, and therefore could not have seeded its own cluster. Without loss of generality, we can assume that the second cluster is C_2 and that its representative sequence is s_2. To define the radius R of the cluster C we need to compute the maximum distance between a sequence within the selected cluster and its representative s_1. Without loss of generality, let us assume that $R_2 = \max(R_1, ..., R_n)$. Given the above, the maximum distance between a sequence within the cluster defined by s_1 and s_1 is the sum of r, the maximum distance between s_1 and s_2, and R_2, the maximum distance between a sequence within the cluster defined by s_2 and its cluster representative, thereby proving the lemma.

Lemma 2: The maximum effective diameter D, is,

$$D = 2 \times \max(R_1, R_2, ..., R_n) + 2r = 2R$$

Proof: Refer to Figure 2.9. The proof follows the same template as that for the radius, except that the selected cluster, C, is assumed to be clustered with an additional cluster, C_3, where $R_3 = R_2 = \max(R_1, ..., Rn)$ and the representative of C_3, s_3, is 2r divergent from s_2.

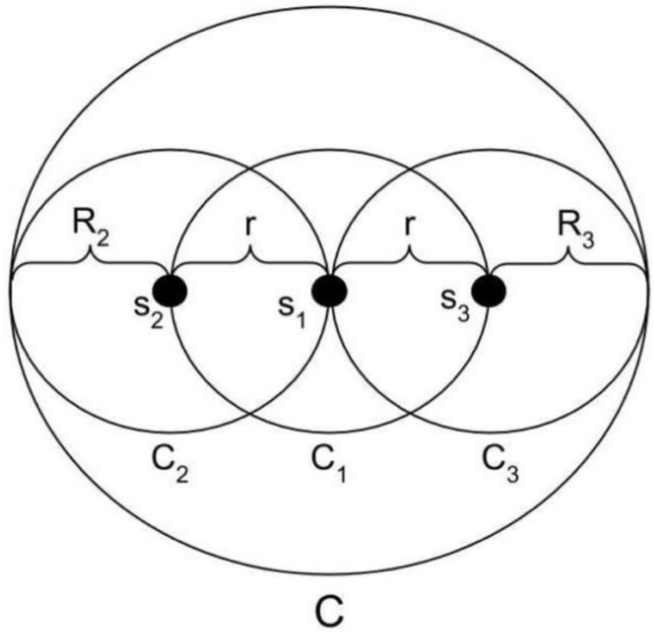

Figure 2.9 Maximum effective radius when clustering multiple representatives
An example cluster, C, with the maximum possible effective diameter when clustering the representatives from C_1, C_2, ..., C_n with tolerance r.

Given the two lemmas, we can now explore the impact on R and D of the number of iterative clustering steps.

2.4.3 Transitive clustering error in the IGC

The diagram in Figure 2.10 highlights the clustering strategy used by the IGC. At each stage, the clustering cut-off, r, was set to 5% divergence (95% identity). Using the formulas derived above, we demonstrate the increase in effective radius and diameter that occurs at each clustering stage, reaching a maximum radius of 15% for the IGC. In other words, within the IGC it is theoretically possible that a sequence may share as little as 85% identity with the corresponding cluster representative, and two sequences that are co-clustered may share as little as 70% identity with each other. These values exceed the nucleotide identity assumed to define a species for the IGC i.e., 95% identity.

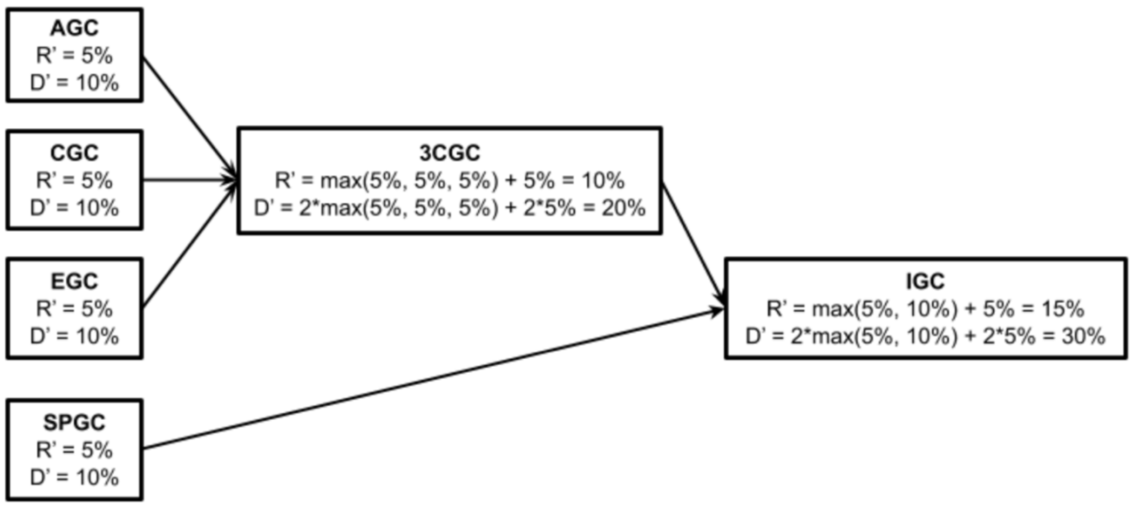

Figure 2.10 Clustering strategy used to create the IGC
Clustering strategy used in creating the IGC. Each block represents catalogs created in the process and shows the worst-case radius and diameter of clusters in the catalog.

2.4.4 Impact of clustering strategy on cluster radius

In Figure 2.11 we compare five different approaches for constructing a catalog from four different catalogs. These range from a single round of clustering that joins all catalogs together in one round (A), to a four-round process that iteratively adds an additional catalog to the previously clustered ones (E). The process used by the IGC is in panel (B). As can be seen in the figure, the final effective radius ranges from 2r (A) to 4r (E), demonstrating how different clustering strategies impact the effective radius of clusters in the final catalog.

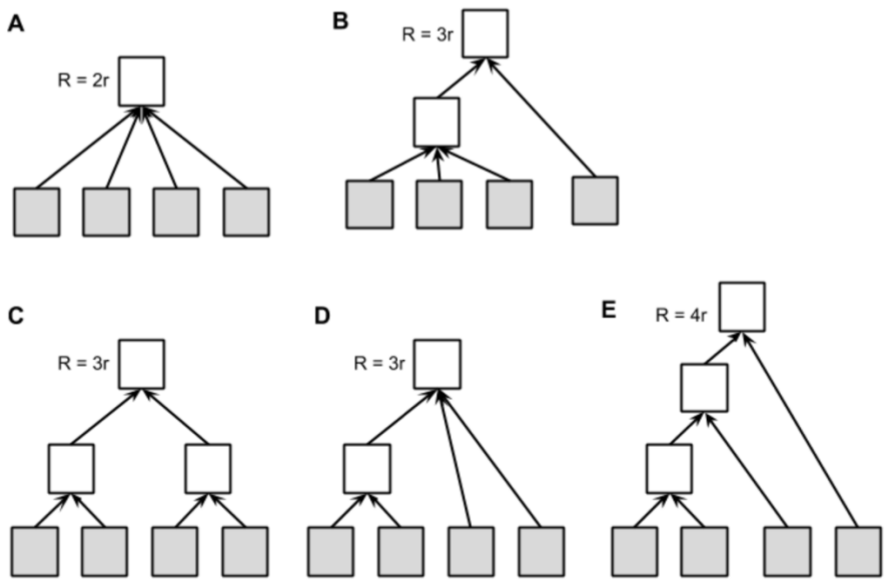

Figure 2.11 Example topologies when combining sequences from four catalogs into one final catalog

Example topologies to combine sequences from four catalogs (grey squares) into one final catalog. The order of combining catalogs can impact the effective radius and diameter of the clusters in the final catalog. For each topology, the effective radius of the final catalog is displayed.

3. Evaluating the accuracy of *Listeria monocytogenes* assemblies from quasimetagenomic samples using long and short reads

This chapter was published in BMC Genomics [39] and was a joint work with Kiran Javkar (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 the bioinformatic analyses, wrote the manuscript and created the figures. AO and PR designed the experiment. AO, PR, YC and NGE performed the microbiological and molecular lab work. ER created the figures and organized the data. NN and DB ran the Opera-MS assemblies and analyses. SC, KJ, AO, PR, YC, NGE, ES, HR, and MP edited the manuscript. All authors read and approved the final version of the manuscript.

3.1 Background

3.1.1 State of the art for pathogen typing

Rapid response, whole-genome sequencing (WGS) networks such as GenomeTrakr [38], PulseNet [110], and the National Antimicrobial Resistance Monitoring System (NARMS) [111, 112] 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 [113]. High resolution MLST analyses involve identifying the profile of alleles for genes in the core genome and whole genome [114, 115], 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 [116-118]. However, despite providing high resolution, SNP and cgMLST/wgMLST analyses do not analyze nor 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 [119].

3.1.2 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% [120], 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 [121]. In contrast, long read sequencing technologies like the Oxford Nanopore platform have higher sequencing error rates (~13% [122, 123]), but can routinely produce reads that are over 10 kbp, thus spanning genomic repeats and supporting the assembly of complete bacterial genomes and plasmids [124].

Although assemblies of nanopore long reads can generate genome-length contigs, they often have many 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 [124]. 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 [125]. Another study reconstructed entire genomes of Shiga-toxin producing *Escherichia coli* strains using nanopore long reads that were polished with Illumina short reads [126]; however, these assemblies had less accurate cgMLST typing compared to those using only MiSeq short reads, despite the short read polishing.

3.1.3 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) [127]. Community dynamics during all types of pathogen enrichments (e.g., *Salmonella enterica*, *Escherichia coli*, *Listeria sp.*) are still poorly understood and co-enriching non-target species often compete with pathogens of clinical significance [128].

3.1.4 Metagenomics

Metagenomics is the direct sequencing of microbial communities [10] 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 [129, 130] contributing complete assemblies for a small subset of the bacteria in the full metagenome [121, 131, 132]. However, achieving sufficient depth of coverage to assemble pathogen genomes directly from metagenomes is often prohibitively expensive.

3.1.5 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 [19, 121]. Quasimetagenomics has been used by FDA scientists since 2009 in efforts to recover pathogens from complex microbiomes such as outbreaks of *Salmonella* in tomatoes [133, 134], to better understand Latin cheese microbiota [135], to look at enrichments for *Salmonella* from cilantro [136], E.coli in flour [137], pathogens in seafood [138-140] and in the public health research response to the Blue Bell ice cream outbreak of 2015, which resulted in the dataset presented here [19, 128]. 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 [19-21]. Many groups are moving the needle forward, demonstrating that strain level differentiation during an outbreak response can be achieved more rapidly with quasimetagenomic approaches [20, 21]. Here we build upon the first ice cream report [128] 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 [19]; and—presented here—the added value of GridIon long reads for circularization of genomes and plasmids.

3.1.6 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 for short read assembly, metaFlye for long read assembly, and Opera-MS for hybrid assembly) as well as popular tools developed for long read genome assembly (Canu and Redbean) and hybrid genome assembly (HybridSpades). Additionally, we evaluated the impact of polishing with three tools: Pilon, ntEdit (both were used to polish long read assemblies with short reads), and Racon (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.

3.2 Results

3.2.1 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, …, B_{30} (Figure 3.1). 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, ..., C_{30}, where $C_n = B_1 + B_2 + ... + B_n$ (Figure 3.1). 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 3.1 records the total number of sequenced bases per C_{30} at each enrichment time.

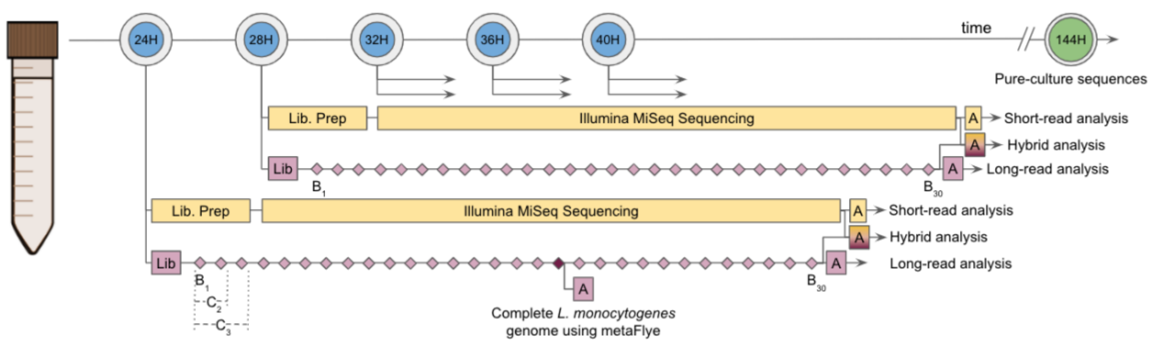

Figure 3.1 The effective time required to sequence and analyze the quasimetagenomic samples
The blue circles marked as 24H, 28H, 32H, 36H, and 40H denote the five enrichment time points where the quasimetagenomic samples were collected and sequenced with the Illumina MiSeq (short read) and the Oxford Nanopore GridIon (long read). Diamonds represent the 30 batches (B_1 to B_{30}) of 4000 GridIon reads, each generated 45 minutes 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., 144 hours. 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.

Table 3.1 Summary of sequence data for C_{30} at each enrichment time.

	24H	28H	32H	36H	40H
Sequenced base pairs	2.3×10^8	3.3×10^8	3.9×10^8	5.4×10^8	5.0×10^8
Number of 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 MiSeq reads sequenced)	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)

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 Nanopore GridIon. The longest sequenced GridIon read was 69,402 nucleotides long (Table 3.2). 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 (Figure 3.2). 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 (Figure 3.3).

Table 3.2 GridIon read length and sequencing error statistics for C_{30}

Enrichment time (hours)	Mean read length	Max. read length	Average quality score	Min. est. sequencing	Max. est. sequencing

				error rate	error rate
24	1923	48588	21.8	7%	18%
28	2721	55258	22.9	6%	17%
32	3268	57233	22.8	7%	16%
36	4445	62426	23.2	6%	13%
40	4129	69402	23.2	6%	13%

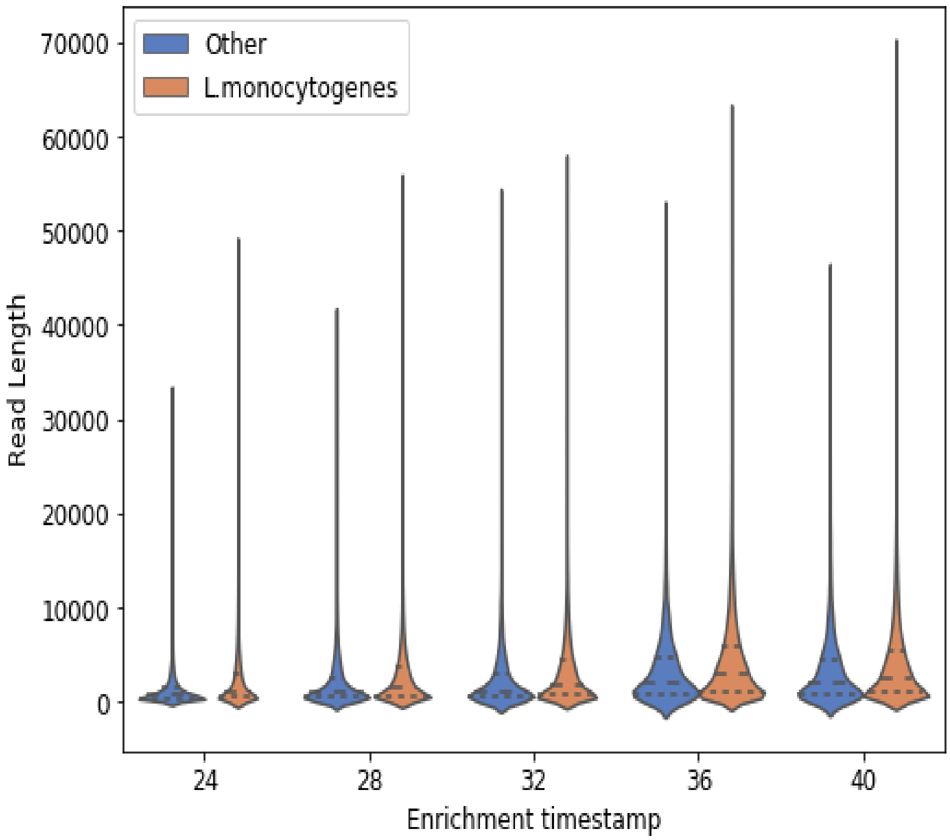

Figure 3.2 Read length distributions for long reads that mapped to the *Listeria monocytogenes* reference genome versus those that did not

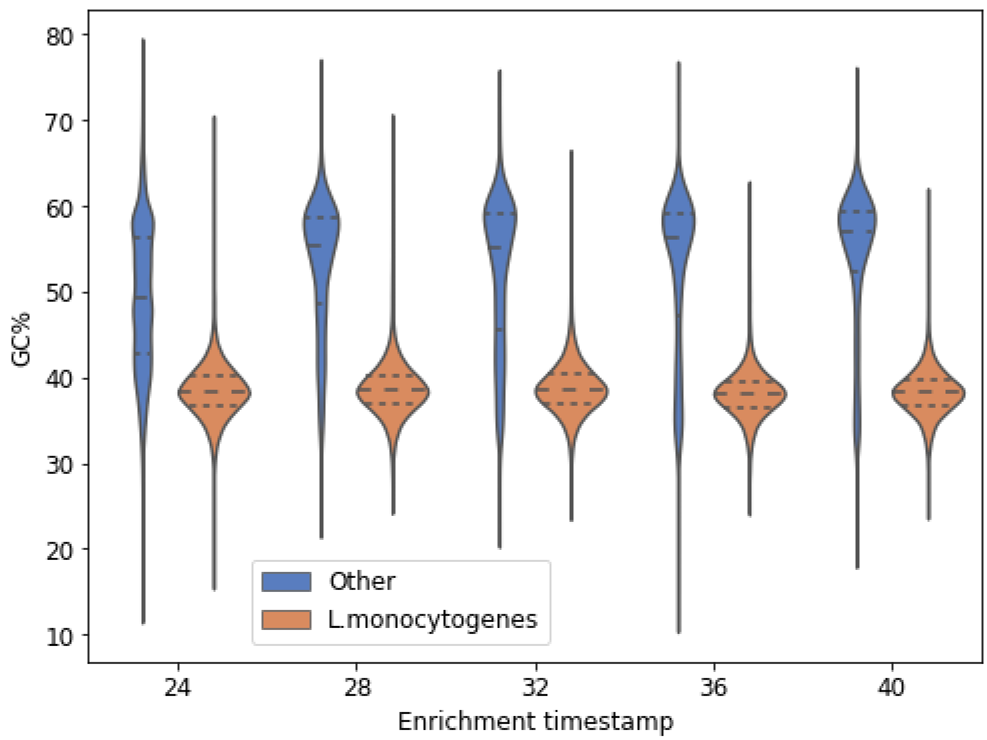

Figure 3.3 GC content distributions for long reads that mapped to the *Listeria monocytogenes* reference genome and those that did not

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 [122]. 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 (18%–7%) to 40H (13%–6.3%) while the mean per-base quality score slightly increased over the same period, from 21.83 to 23.19, respectively.

3.2.2 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 [141]. The reference was treated as a "gold standard" with an expected accuracy of ~99.999% [142]. Previous research had shown that the outbreak consisted of two strains. One that was only isolated from Facility 1 and another that was mainly isolated from Facility 2 [141]. 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 [143] distance: 0.0206 and 0.0218 respectively). The reference from Facility 1 was subsequently used for our analysis.

The similarity between the *L. monocytogenes* contigs derived from the quasimetagenomes and the reference sequence was assessed, and 55 loci were identified (46 single nucleotide insertions, 2 di-nucleotide insertions, and 5 single nucleotide polymorphisms) that differed at all enrichment times. Four of these variants (1 single nucleotide polymorphism 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).

3.2.3 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 [128] 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.

3.2.4 General quasimetagenome assembly statistics

Ten assembly approaches were tested (Table 3.3), 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).

Table 3.3 The ten assembly approaches tested

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 (1) de novo or (2) using reference genomes	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

All assembly approaches had a mean runtime (for the full set of reads, C_{30}, across enrichment times) of approximately 40 minutes or less (Table 3.4) except Canu which had a mean runtime of 98 minutes per sample. The fastest assembly approach was Redbean with a mean runtime of just one minute (Figure 3.4).

Table 3.4 Mean assembly statistics (C_{30} at each enrichment time) for each assembly approach

Assembly tool	Runtime	Total assembly length	Number of contigs	N50	Longest contig
metaFlye (long read)	40.6	4291417	27	3056133	3056133
Canu (long read)	98	3470967	21	1754979	2071553
Redbean (long read)	1	3474503	35	2123769	2131343
MegaHit (short read)	32.8	7972605	7315	97577	672182

metaFlye+Racon (long read)	41.6	4261624	27	3039238	3039238
HybridSpades (short read hybrid)	22.6	11681048	19285	112850	686270
OperaMS (no reference) (short read hybrid)	12.2	10340211	13921	105382	655220
OperaMS (reference) (short read hybrid)	13.6	10363273	13913	205943	1919416
metaFlye+Racon+Pilon (long read hybrid)	41.6	4271759	27	3041086	3041086
metaFlye+Racon+ntEdit (long read hybrid)	41.6	4274358	27	3041440	3041440

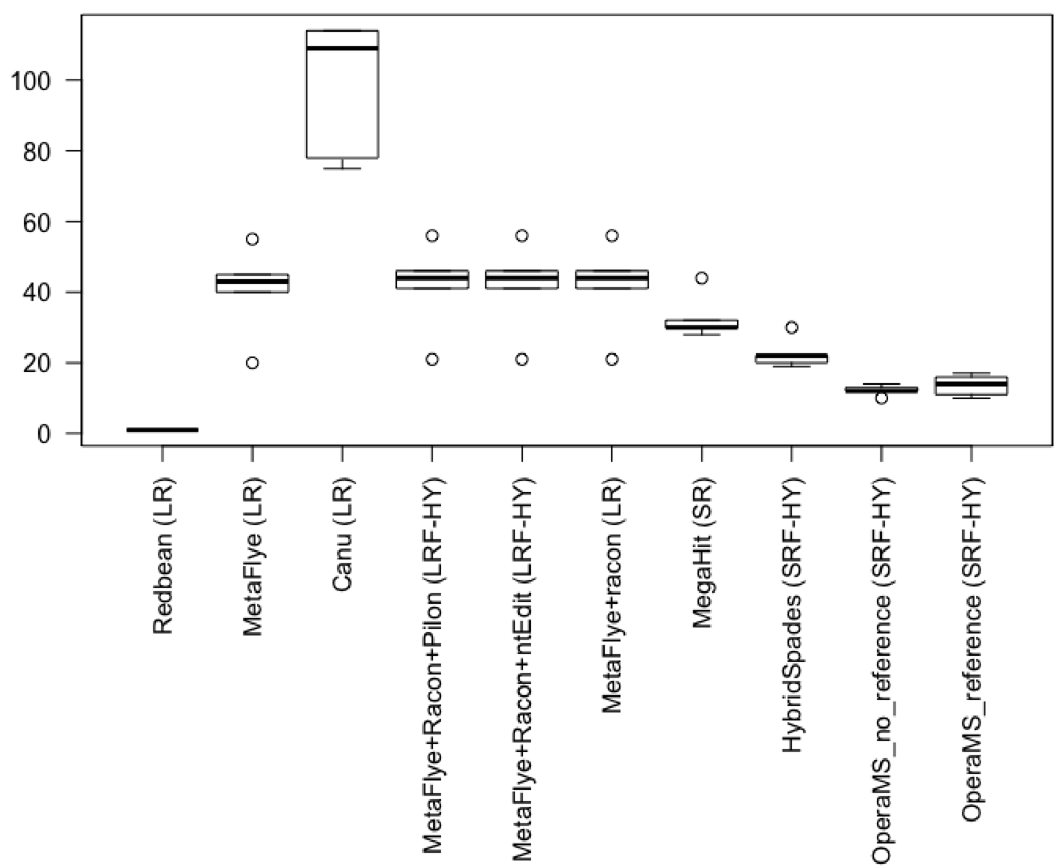

Figure 3.4 Runtimes for the assembly approaches

Runtimes for the assembly approaches in minutes when assembling C30 at each enrichment time. (Abbreviations: SR = short read, LR = long read, HY = hybrid).

The contiguity of the assemblies was measured using several metrics: the total assembly length (Figure 3.5), number of contigs (Figure 3.6), N50 (Figure 3.7), and longest contig assembled (Figure 3.8).

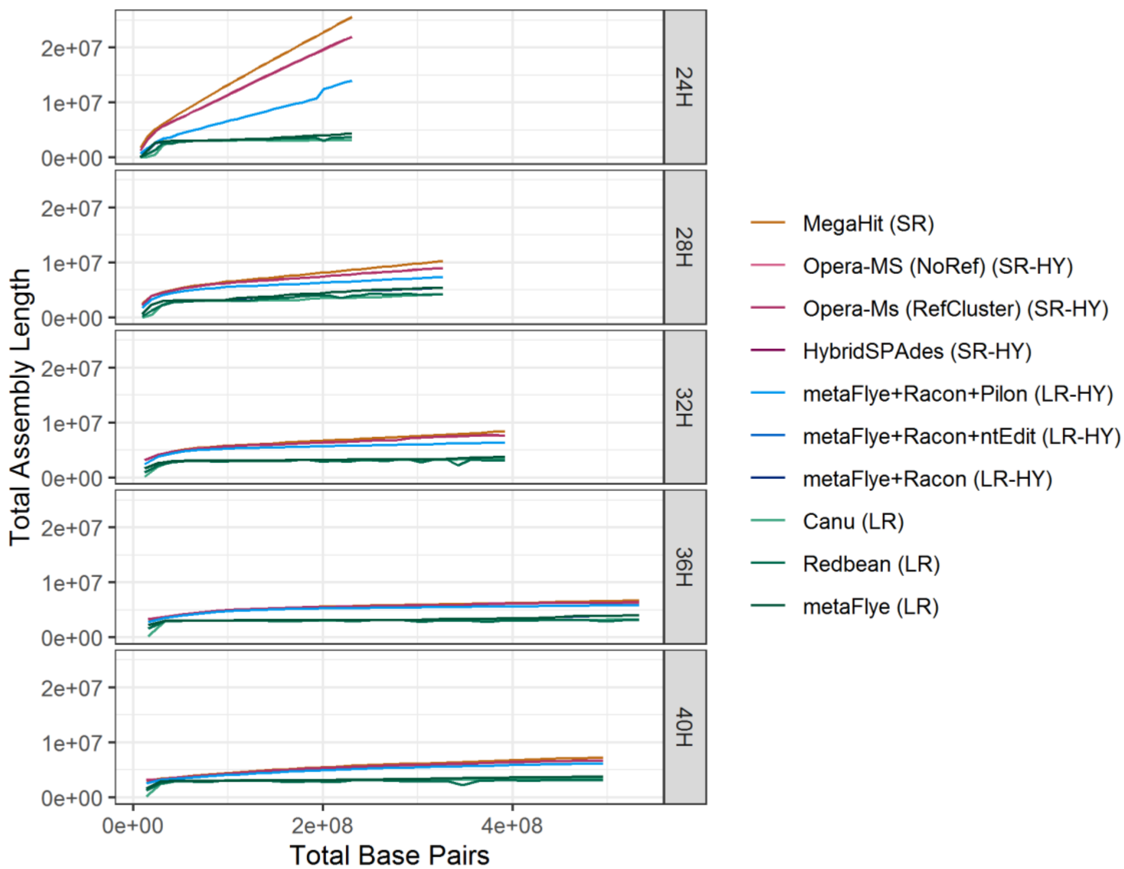

Figure 3.5 Total assembly length per base pairs sequenced for each assembly approach
Total assembly length versus the total number of base pairs sequenced per cumulative batch at each of the enrichment time points for each assembly approach. Sometimes the results for Canu, Redbean and metaFlye overlap as do Opera-MS (NoRef) and Opera-Ms (RefCluster). (Abbreviations: SR = short read, LR = long read, HY = hybrid)

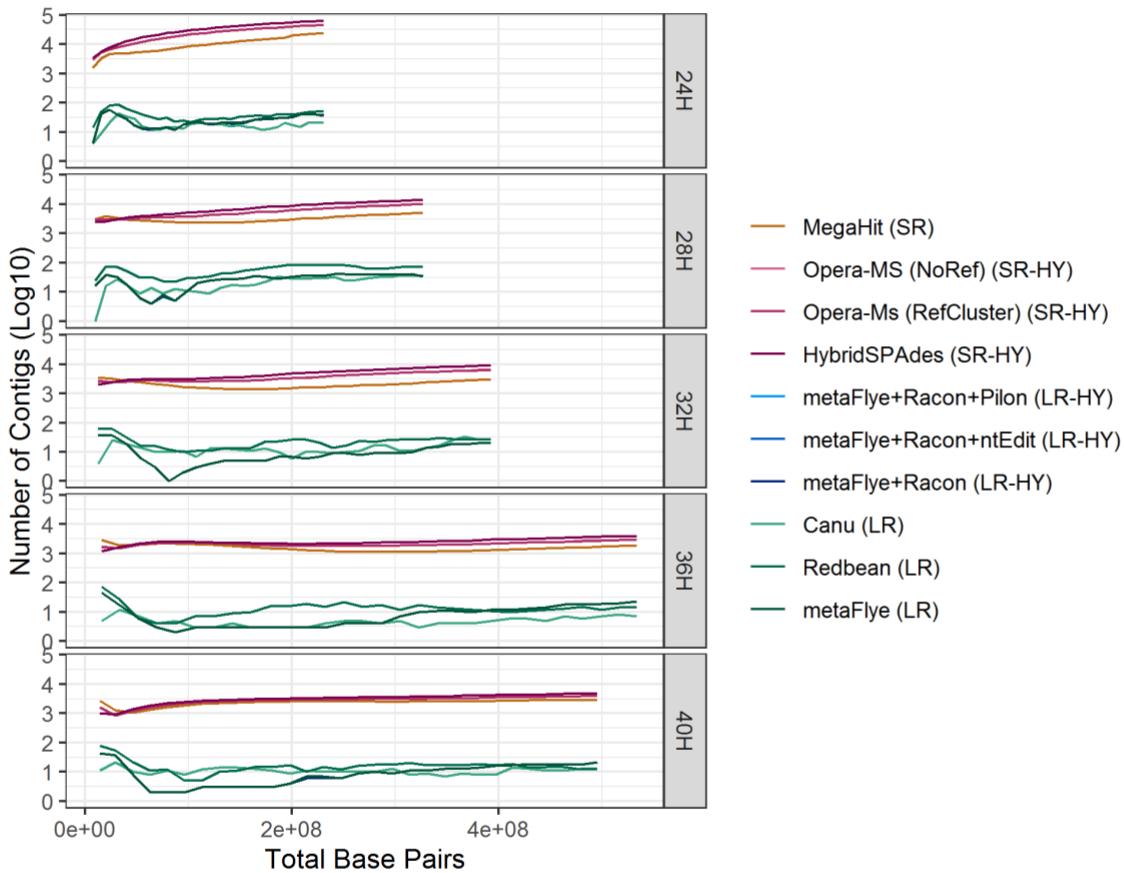

Figure 3.6 Number of contigs per base pairs sequenced for each assembly approach

Number of contigs versus the total number of base pairs sequenced per cumulative batch at each of the enrichment time points for each assembly approach. metaFlye+Racon, metaFlye+Racon+Pilon, and metaFlye+Racon+ntEdit are obscured by the line for metaFlye in each of the plots. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

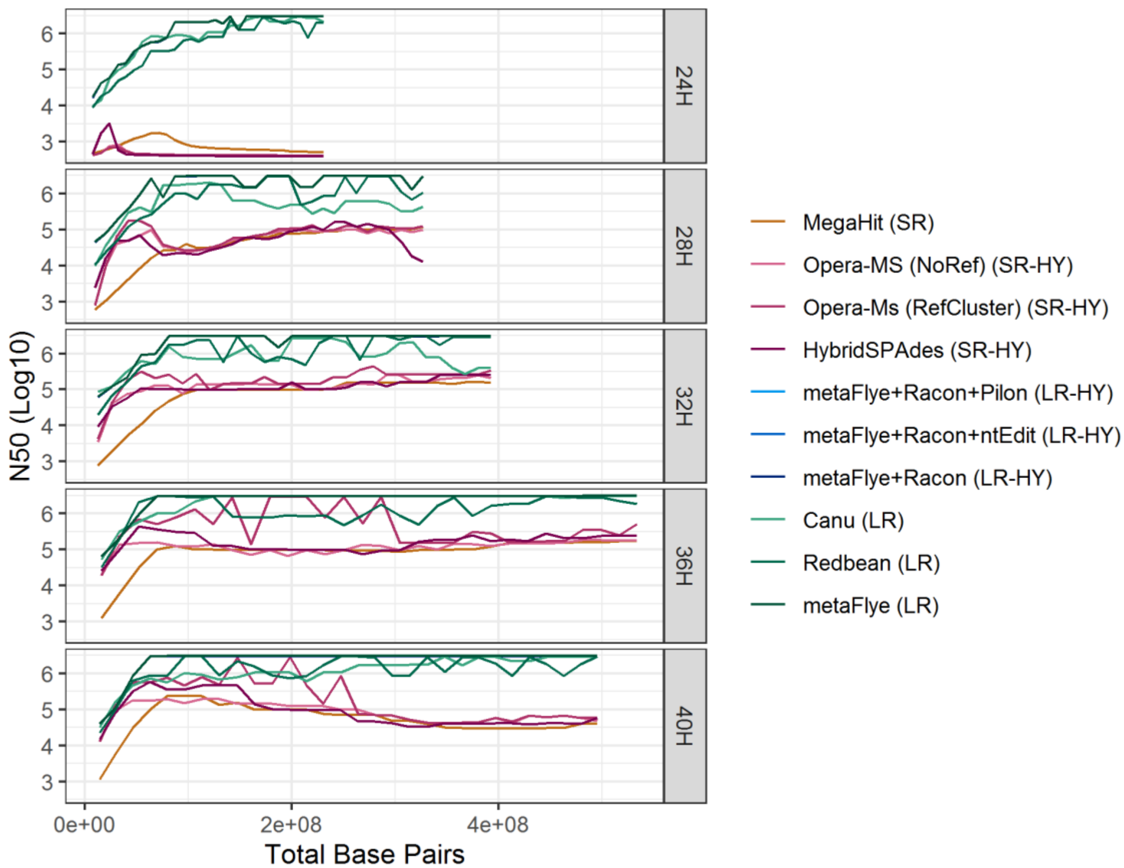

Figure 3.7 N50 per base pairs sequenced for each assembly approach

N50 versus the total number of base pairs sequenced per cumulative batch at each of the enrichment time points for each assembly approach. metaFlye+Racon, metaFlye+Racon+Pilon, and metaFlye+Racon+ntEdit are obscured by the line for metaFlye in each of the plots. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

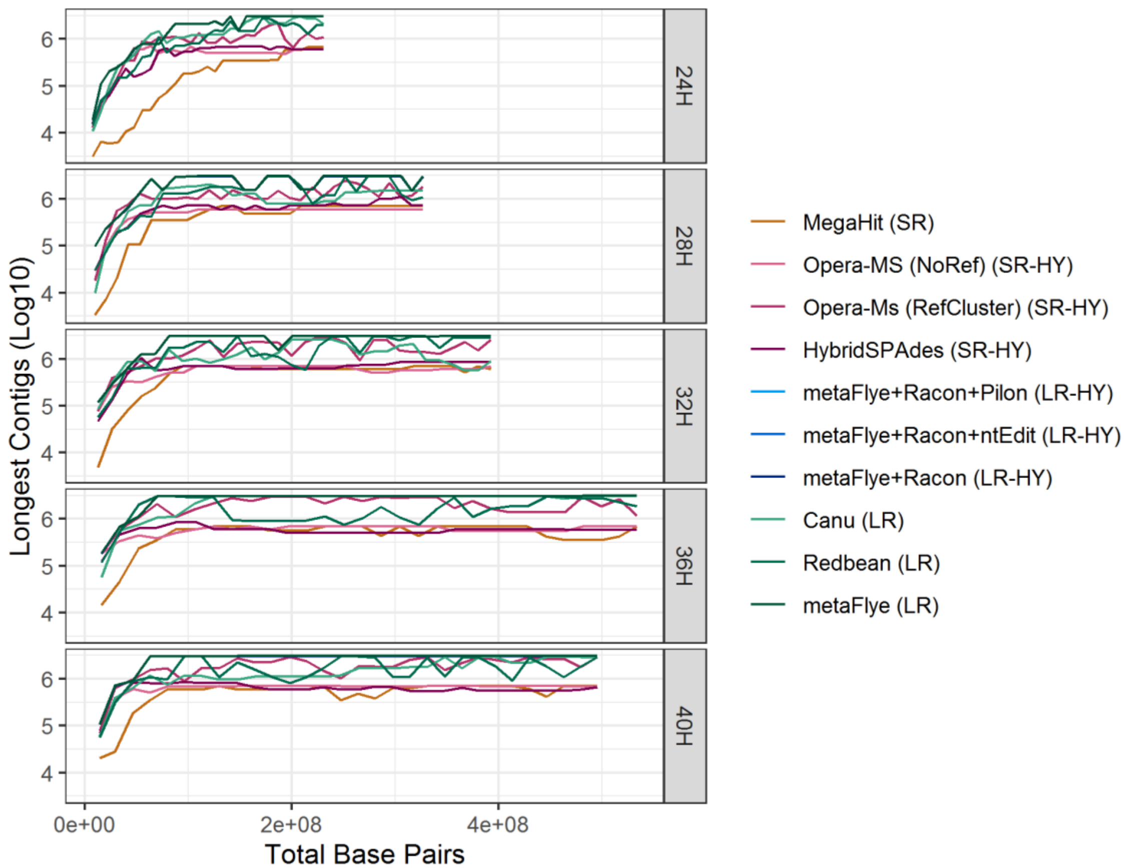

Figure 3.8 Longest contig assembled per base pairs sequenced for each assembly approach
The longest contig assembled versus the total number of base pairs sequenced per cumulative batch at each of the enrichment time points for each assembly approach. metaFlye+Racon, metaFlye+Racon+Pilon, and metaFlye+Racon+ntEdit are obscured by the line for metaFlye in each of the plots. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

The mean values for C_{30} across enrichment times for each contiguity metric are described in Table 3.4. 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 (Figures 3.5 and 3.6).

As expected, the long read and long read hybrid assemblies had the highest 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 the highest N50 values with the longest contigs nearest to the length of the *L. monocytogenes* reference genome (Table 3.4); 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 60X 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.

3.2.5 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 3.9). 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*).

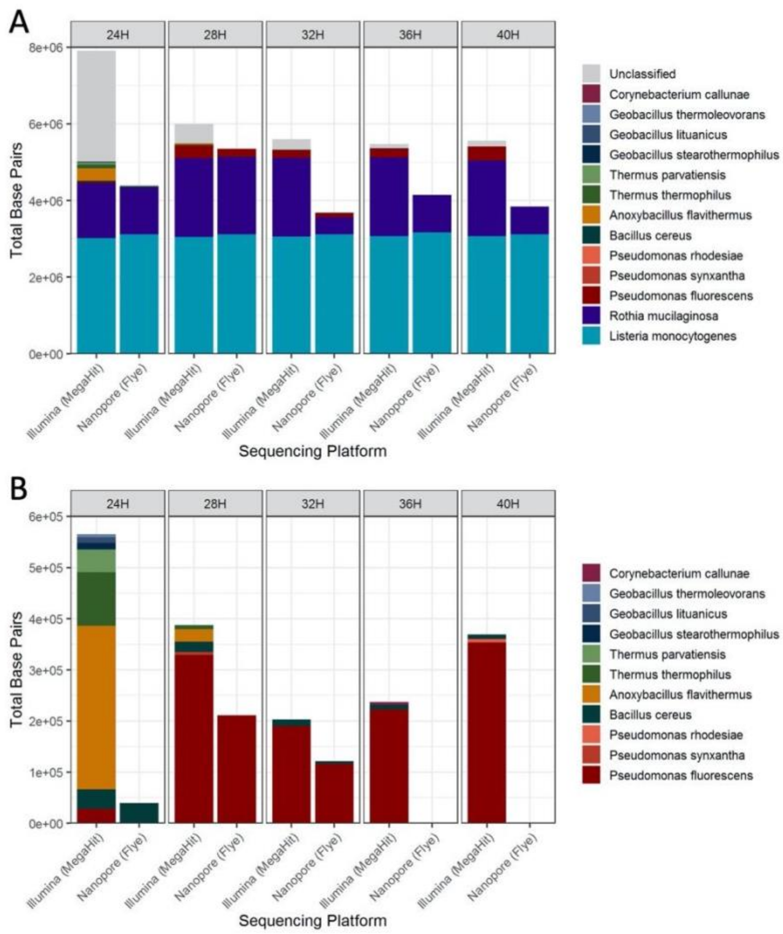

Figure 3.9 Taxonomic composition at each enrichment time

Taxonomic classification of cumulative batch 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 5,000 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 always the most abundant species and its abundance increased with enrichment time, but the abundance estimates differed for the MiSeq and GridIon (Table 3.5). At 24H, 33% and 60% of the MiSeq and GridIon reads, respectively, mapped to the *L. monocytogenes* reference genome. At 40H, 92% and 97% of the MiSeq and GridIon reads respectively, mapped to the reference genome.

Table 3.5 Percent of reads that map to the *L. monocytogenes* reference genome.

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

3.2.6 Reconstruction of the *L. monocytogenes* genome from the 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 3.10). 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. Because metaFlye assembled genome-length contigs for *L. monocytogenes* the most consistently of the long-read assemblers, only the metaFlye assemblies were used for the 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.

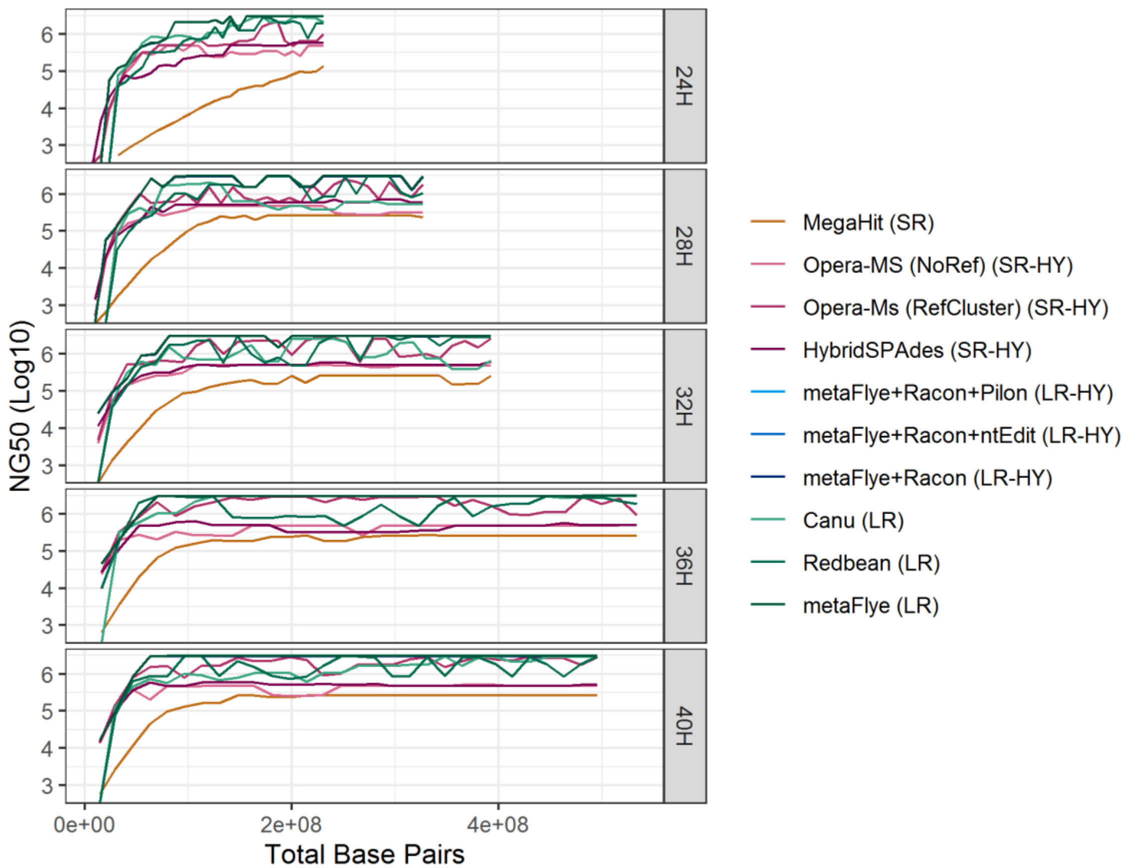

Figure 3.10 NG50 per base pairs sequenced for each assembly approach
The NG50 versus the total number of base pairs sequenced per cumulative batch for the assembled *L. monocytogenes* contigs at each of the enrichment time points for each assembly approach. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

The 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 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 (33X depth of coverage of the *L. monocytogenes* genome), 24H and C_{22} with Canu (40X depth of coverage of the *L. monocytogenes* genome), and 28H and C_{16} with Redbean (47X 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.

3.2.7 Assembly errors in the *L. monocytogenes* genomes reconstructed from the 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 3.11). 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 ~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.

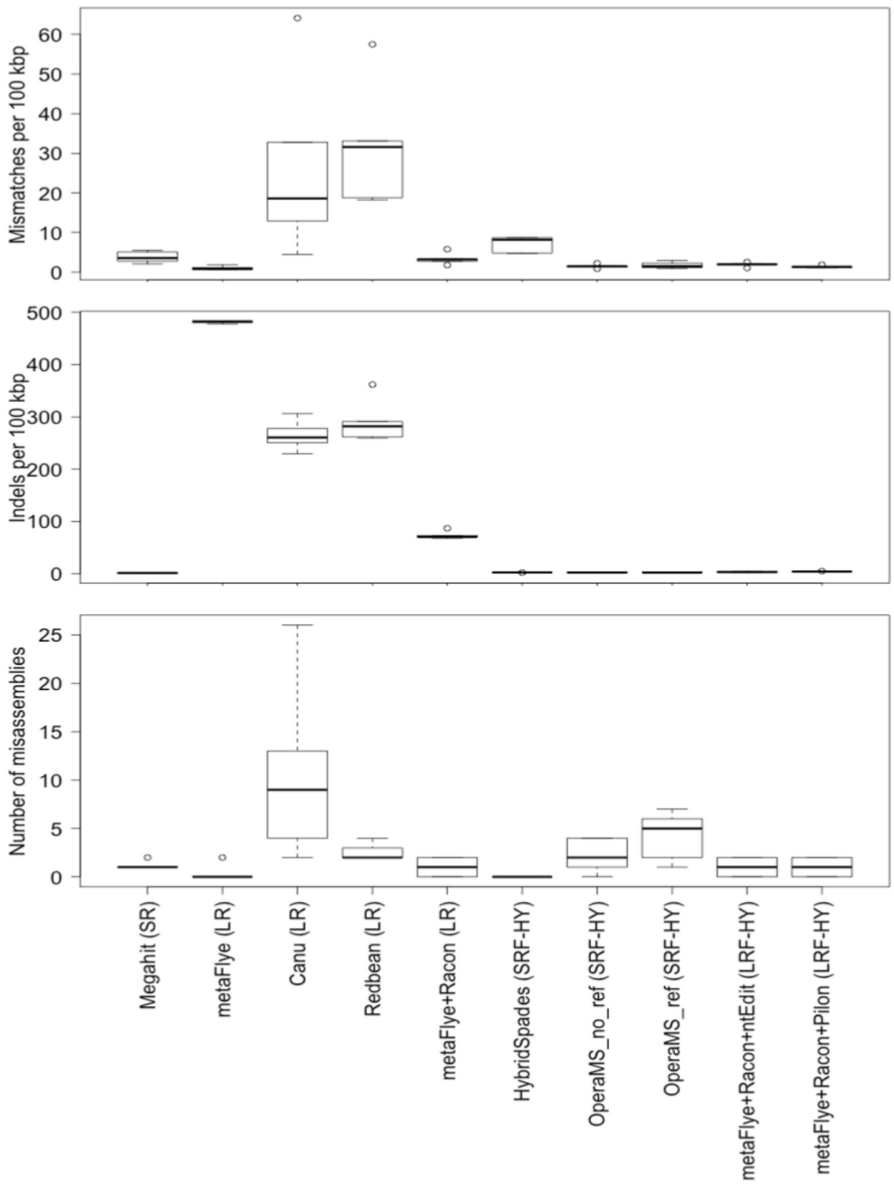

Figure 3.11 The quality of assembled contigs annotated as *L. monocytogenes*, with respect to the reference genome

The quality of assembled contigs annotated as *L. monocytogenes*, with respect to the reference genome, using Quast for cumulative batch 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)

3.2.8 Accuracy of the *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 ~160X 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 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_1 to C_{30} at each enrichment time) was assessed by calculating the BLAST distance between the core genes (Figure 3.12) and the complete set of genes (Figure 3.13) of the reference genome and the *L. monocytogenes* contigs. As defined earlier, the BLAST distance is a measure of sequence similarity equaling 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.

For the core genes, the smallest BLAST distance observed was 5 (Figure 3.12). 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.

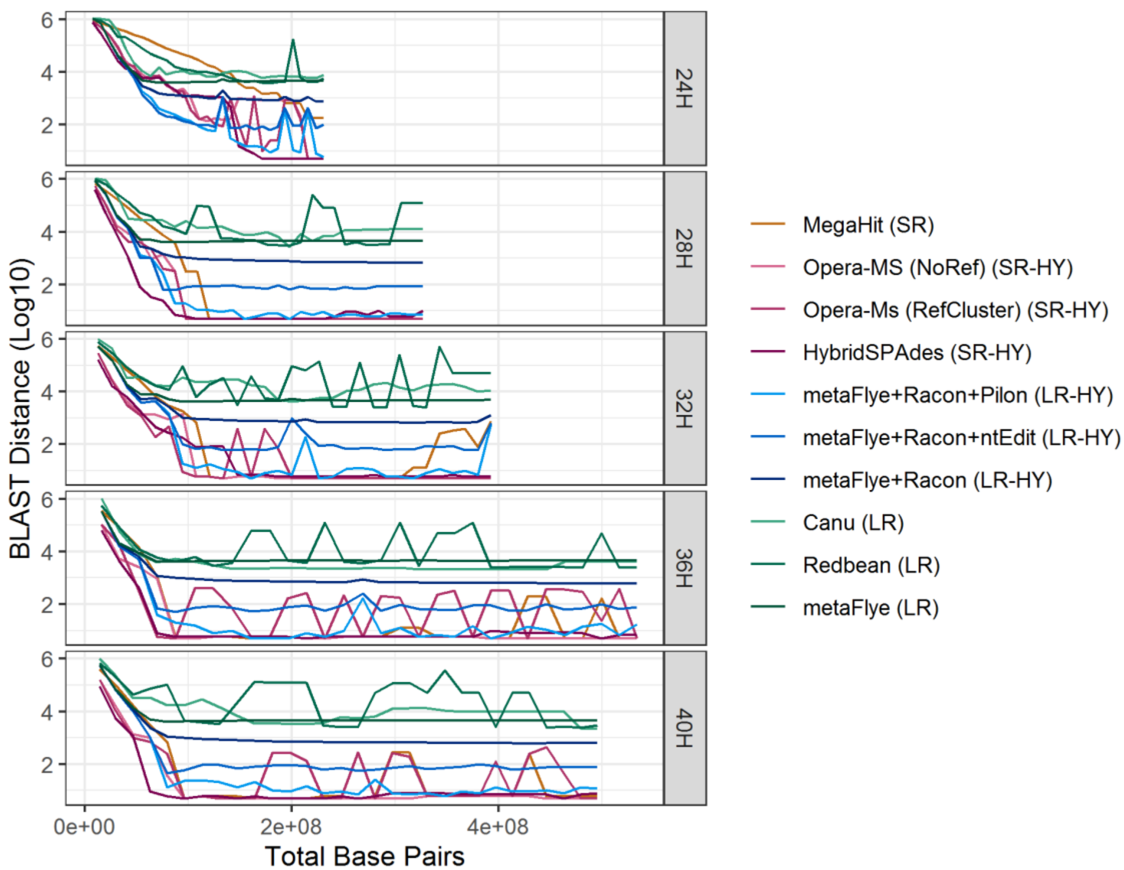

Figure 3.12 Core gene BLAST distances
BLAST distance between the core genes of the reference genome and the assemblies versus the total number of base pairs sequenced per cumulative batch. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

The 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 40X (long reads) and 19X (short reads) depth of coverage of the *L. monocytogenes* reference genome; Opera-MS, both with and without reference-guided scaffolding, at 24H and C_{28} corresponding to 50X (long reads) and 25X (short reads) depth of coverage of the *L. monocytogenes* reference genome. Megahit assemblies attained a BLAST distance of 5 after 28H and C_{11} corresponding to 28X depth of coverage of the *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 2,000 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 the *L. monocytogenes* reference genome; however, it achieved BLAST distance 5 less consistently than short read or short read hybrid approaches (Figure 3.12). Long read hybrid assembly with ntEdit assembled the core genes with less accuracy than Pilon, with a median BLAST distance (C_1 to 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 3.13). Pilon achieved a BLAST distance of 132, the best observed for any tool, at 28H and C_{14} corresponding to 36X (short reads) and 38X (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 2,991.

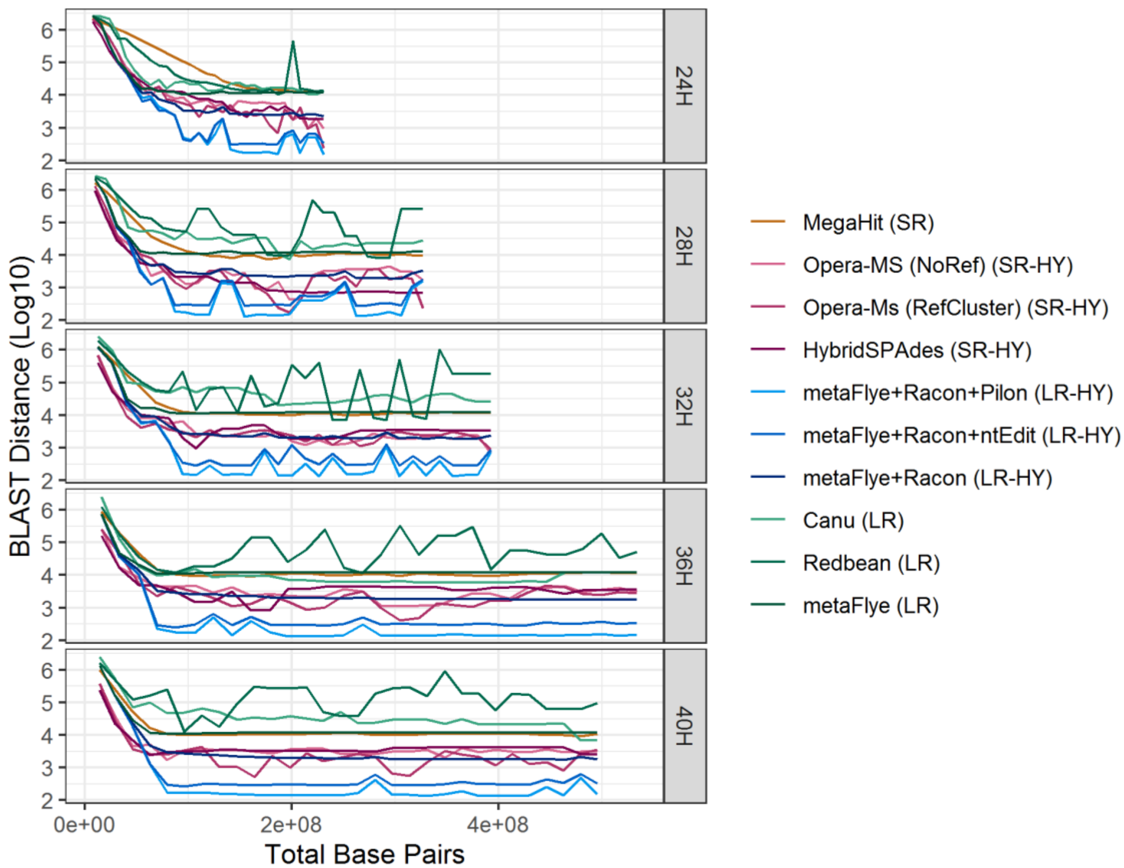

Figure 3.13 Complete gene set BLAST distances
BLAST distance between the complete gene set of the reference genome and the assemblies versus the total number of base pairs sequenced per cumulative batch. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

3.2.9 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 the 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 3.14).

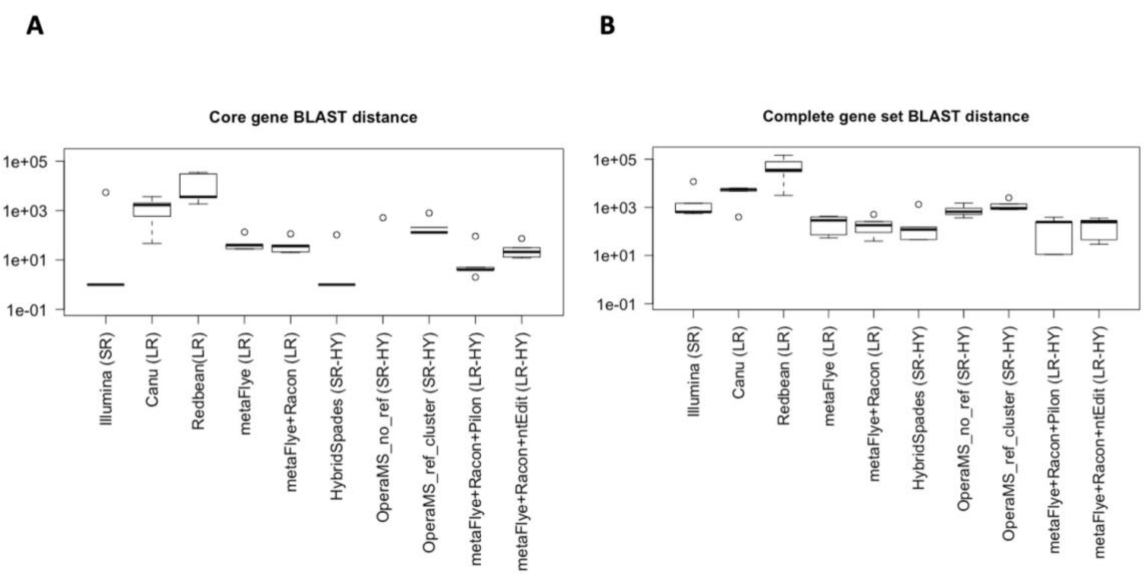

Figure 3.14 Consistency of assembly approaches between successive cumulative batches
Median successive cumulative batch difference in BLAST distances, across enrichment time points, for the A) the core genes and B) the complete gene.

The tools that most consistently assembled the core genes were Opera-MS (without reference guided scaffolding), HybridSpades, MegaHit, and metaFlye+Racon+Pilon; 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 (1,183) 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 consistent 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 1,000, with the exceptions of Opera-MS with reference guided scaffolding (1,019), Canu (4,758), and Redbean (32,655).

3.2.10 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 30X depth of coverage and at 150X depth of coverage, 695,760 nt and 695,778 nt, respectively. 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 100X depth of coverage of the *L. monocytogenes* genome for both short and long reads.

3.3 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 SNPs 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.

3.3.1 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 [19-21]. 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 24 hours enrichment—a significant improvement over the ~6 day procedure required to culture and sequence an isolate genome [19]. 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 28 hours 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 24 hours of enrichment (Figure 3.1). This highlights that an integrated approach to quasimetagenomics can greatly expedite and enhance source tracking.

3.3.2 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 [121]. 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 ~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 24 hours of enrichment and only 33X 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 mobilome. However, similar to previous studies [144], 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 minutes (as opposed to a MiSeq sequencing run which takes ~24 hours depending on the number of cycles). As assembling the reads is much faster than sequencing itself (Figure 3.1), 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.

3.3.3 Hybrid assembly outperforms other approaches but with trade-offs

Our results support the accuracy of hybrid assemblies[20, 125, 126]—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 have 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.

3.3.4 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 preferrable with the added value of assembling complete genomes and mobile elements like plasmids.

3.3.5 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.

3.4 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 24 hours enrichment with the short read hybrid assemblies and 28 hours 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.

3.5 Methods

3.5.1 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 3.1). The isolation of a pure colony of *L. monocytogenes* for sequencing typically requires up to 6 days of selective culture enrichment [19]. 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.

3.5.2 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 [127]. 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 acriflavine, 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 40-hour enrichment.

3.5.3 DNA extraction and sequencing for short reads

For each enrichment time point (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 200mL of enzymatic lysis buffer containing 20mM Tris-HCl (pH-8.0), 2mM Sodium EDTA, 1.2% Triton X- 100, 20mg/ml of lysozyme. Samples were incubated for 60 min at 37 ºC. 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 x 250 cartridge (Illumina, CA) following manufacturer recommended protocol.

3.5.4 DNA extraction and sequencing for long reads

For each enrichment time point (24H, 28H, 32H, 36H, 40H), 2ml 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 300ul of TE Buffer. 300ul 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 4,000 sequenced reads in fast5 format files [145]. Each fast5 file was converted into fastq formatted DNA sequences using Guppy for base calling [146]. The fast base calling mode was used, which has a speed of ~4.6 Mbp/second. The GridIon typically outputs a batch of reads every 30 to 60 minutes (internal to lab) but is affected by factors such as the length and quality of the DNA fragments being sequenced.

3.5.5 *L. monocytogenes* reference genome

Previous work identified two strains from the *L. monocytogenes* ice cream outbreak [141]. 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 [141]. The reference is 3,030,827 bp long with 2,984 protein-coding genes (2,710,041 bp in total length) predicted by Prokka (v1.12) [93] and a GC-content of 38%. The core genes (using the 1,013 gene cgMLST scheme developed for *L. monocytogenes* at the FDA [114, 115]) were identified by BLAST [84] 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 [147].

3.5.6 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 nth batch. Our analysis used the first 30 batches of reads, i.e., the first 120,000 reads corresponding to batches B_1, B_2, …, B_{30} (Figure 3.1). 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, ..., C_{30}, where $C_n = B_1 + B_2 + ... + B_n$ (Figure 3.1). 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 1 records the total number of sequenced bases per C_{30} at each enrichment time.

3.5.7 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) [148] if there was ≥10X depth of coverage and ≥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) [77] and analyzing the pile-up of reads with SAMtools (v1.7) [149]. Loci with ≥50X 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 treated as evidence for multiple strains.

3.5.8 Raw read statistics and reference genome coverage

Raw read statistics were collected for the 30 batches of reads (B_1 to 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) [150], 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.

3.5.9 Assembling the sequenced reads

Short reads and long-reads from each cumulative batch (C_1 to C_{30}) were assembled per enrichment time point (Figure 3.1). The short reads were assembled using MegaHit (v1.2.9) [92] with default settings and scaffolded with MetaCarvel [151]. The long reads were assembled using Canu (v1.7) [152], Redbean (v2.5) [153], and metaFlye (v2.6-release) [154] 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 the metaFlye assemblies were polished, using the long reads, with Racon (v1.4.15) [155]. HybridSpades (v3.14.0) [156] and Opera-MS (v0.8.3) [144] (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) [157] and ntEdit (v1.3.1) [158] 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.

3.5.10 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) [159] 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 to C_{30} at each enrichment time) using a custom Python script.

3.5.11 Comparison of the reference genome with the *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 [160].

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 to 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 sequenced assembled from a sample comprised a single contig, the synteny of the core genes was compared to that in the reference.

3.5.12 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) [161] and the MiniKraken database using default settings. A species was considered present if ≥5,000 nt 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.

4. taxaTarget: Fast, sensitive, and precise classification of microeukaryotes in metagenomic data

This chapter has been submitted to BMC Microbiome (available as a preprint [162]) where it is currently under peer review. It was a joint work with Kiran Javkar, Harihara Subrahmaniam Muralidharan, Padmini Ramachandran, Andrea Ottesen, Hugh Rand, and Mihai Pop. SC, AO, MP, and HR designed the study. SC, MP, HR, KJ, and HM designed the algorithms. SC analyzed the data and wrote the manuscript. SC, MP, HR, KJ, and AO interpreted the data. All authors read, provided feedback, and approved the final manuscript.

4.1 Background

Eukaryotes, bacteria, and archaea comprise the domains of life on Earth (according to our current understanding). Well known eukaryotes, like plants and animals, belong to two small branches on a much larger phylogenetic tree of mostly uncharacterized microscopic eukaryotes, i.e., microeukaryotes. Some microeukaryotes are well known because they benefit humans. For example, *Saccharomyces cerevisiae* (baker's yeast) and *Aspergillus sojae* are involved in the creation of fermented foods, drinks, and spices such as beer, bread, miso, and soy sauce. Others, like *Plasmodium falciparum*, *Trypanosoma brucei*, and *Cyclospora cayetanensis* are infamous for causing human diseases like malaria, sleeping sickness and food poisoning outbreaks, respectively. However, studies estimate that less than 5% of the predicted millions of species of microeukaryotes are currently characterized [163-165]. Furthermore, the eukaryotic tree of life is still expanding as whole new supergroups are being discovered [166].

Microeukaryotes do not exist in isolation, rather they are ecologically integral members of diverse microbial communities (alongside bacteria, archaea, viruses) i.e., microbiomes, found nearly everywhere on Earth. For example, phytoplankton, a mixture of photosynthetic microorganisms floating in oceans and freshwater bodies, are responsible for nearly half of all carbon fixation and photosynthetic primary production on Earth—70% is performed by microeukaryotic groups i.e., Diatoms, Coccolithophores, and Chlorophytes [167]. In the human gut microbiome, some species (e.g., *Blastocystis* and *Entamoeba*), formerly considered strict parasites, might promote healthy and diverse assemblages of bacterial species and only express pathogenicity in certain microbial contexts [168-173].

A useful tool for exploring the diversity of microeukaryotes is metagenomics. Metagenomics is the sequencing of DNA fragments, referred to as reads, extracted directly from a sample. Such data provides a means for studying the diversity, functional potential, and ecology of microbiomes. Although metagenomics has led to an explosion of knowledge about microbial communities [174], most studies have focused on bacteria, with relatively few (although a growing number) focusing on microeukaryotic groups such as fungi and protists [171, 175-177].

The metagenomic study of microeukaryotes has lagged for a variety of reasons. Perhaps most important, is that there is a general sparsity of sequenced genomes and transcriptomes, with more eukaryotic species represented in protein databases, e.g., ~40,000 species in the UniRef100 database, than genome databases, e.g., ~4,000 species in the National Center for Biotechnology information (NCBI) Assembly database. Combined, this is still orders of magnitude fewer species than the millions estimated to exist [163, 164]. Sparse reference sequences can negatively impact the performance of tools that use sequence similarity for taxonomic classification. For example, metagenomic reads from a novel species might not align to any sequences in the reference database of a tool and thus go undetected.

Another challenge for metagenomics is that microeukaryotes often occur at low abundance in microbial communities with genomes that are orders of magnitude larger than bacteria [178, 179]. Ideally, all the genomes in a metagenome could be directly reconstructed (i.e., assembled) from the reads. However, the limitations of current sequencing technologies, and the general sparsity of eukaryotic reads per metagenome present algorithmic challenges for assembly that often result in few or no overlapping reads to construct longer eukaryotic sequences (i.e., contigs). Microeukaryotic contigs longer than 3,000 bp can sometimes be assembled, given sufficient depth of coverage [180], and classified as eukaryotic based upon k-mer composition with high accuracy [181]. However, current bioinformatic approaches for characterizing microeukaryotes tend to rely upon metagenomic reads for analysis.

An alternative to the classification of assembled contigs is mapping the reads to a reference database of sequences (e.g., genomes or genes). However, for large eukaryotic genomes (sometimes billions of base pairs), mapping metagenomic reads directly can be computationally expensive in terms of runtime and memory. Additionally, the assemblies of eukaryotic genomes in reference databases are frequently contaminated with bacteria, potentially leading to classification errors [182-184]. A more robust and scalable solution is the use of marker sequence databases which can be clade-specific (as used in MetaPhlan3 [185]) or universally present such as ribosomal subunits (as used by Metaxa2 [186] and mTAGs [187]) or single copy orthologs (as used by EukDetect [184, 188]). The latter have shown high specificity for eukaryotes with a strong phylogenetic signal [180, 188]. A drawback for using a database of marker sequences is that the combined length of the marker sequences is usually a small fraction of any eukaryotic genome, potentially resulting in low sensitivity when employed in metagenomic data.

A general challenge for tools using sequence similarity for taxonomic classification is that sequences, and regions of sequences, can have variable levels of discriminatory power because they evolve at different rates [189]. For example, regions of the 18S ribosomal subunit (rRNA), a eukaryotic marker sequence used by Metaxa2 and mTAGs for taxonomic classification, are known to lack discriminatory power at the species, genus, and sometimes even the family or broader levels for some eukaryotic groups [165]. One consequence is that a metagenomic read can align equally well to a sequence in multiple species, providing insufficient evidence for classification at the species level. EukDetect addressed this problem by removing marker genes from its database to which bacterial reads mapped, or if the marker gene Hidden Markov Models (HMM) misclassified bacterial sequences as eukaryotic [184]. This approach, which was not extended to other groups like archaea or viruses, reduced the number of marker gene families used for taxonomic classification by EukDetect from 255 to 214. One drawback of discarding entire eukaryotic marker gene families with high similarity regions in bacteria is that unshared regions of the marker genes, which might be informative for eukaryotic taxonomic classification, also get discarded.

Another issue for approaches using sequence similarity for taxonomic classification is the application of universal alignment score cutoffs for classifying metagenomic reads that map to database sequences. For example, reads might be classified as the species of the database sequences they align to if the alignment is ≥95% identity. This could be extended to genus if the alignment is ≥85% identity, family if the alignment is ≥75% identity, and so on. EukDetect and Metaxa2 use similar strategies [184, 186], but such methods ignore that sequences can evolve at different rates and can have variable discriminatory power. A more sophisticated approach, used by CCMetagen, learned the alignment score cutoffs, from phylum to species, using the sequence similarity and taxonomic relatedness of fungal marker sequences [190]. Although a data-driven approach, the thresholds learned for fungal sequences end up being applied to the entire NCBI Nucleotide database which contains sequences across the tree of life and viruses. Another approach for the taxonomic classification of metagenomic reads, used by mTAGs [187], is to assign the lowest common ancestor of the database sequences each read aligns to. However, such approaches tend to classify reads with too high of specificity when using sparse databases (such as exists for microeukaryotes) and to be too conservative when using highly representative databases [104, 191].

As an alternative, supervised learning approaches have been developed for dynamically determining thresholds using marker genes [192] and for the discovery of protein families [193] in metagenomic data; however, these approaches were developed for bacteria using bacterial marker genes or for functional profiling instead of taxonomic classification. Supervised learning methods use labeled training data to learn functions for classifying new examples. For example, MetaPhyler [192] predicts the taxonomic relatedness (from phylum to species) of metagenomic reads aligned to marker genes. It learns the cutoffs for classification thresholds by training on the scores of reads, from the genomes of known taxa, aligned to the marker genes in its database. This approach automatically assesses if sequences contain enough information for classification at different taxonomic levels. However, a limitation of MetaPhyler is that it builds classifiers for whole genes, treating all regions of a gene as having equal classification power. ROCker [193] introduced a method for assessing if metagenomic reads were sequenced from specific gene families using sliding window, region-specific classifiers. Although a functional profiling tool, we thought the approach could be successfully adapted for taxonomic classification.

Here, we present a metagenomic taxonomic profiling tool, taxaTarget, specifically designed for microeukaryotes. We show that taxaTarget is computationally efficient, with higher sensitivity, and often higher precision than similar tools using marker sequence databases—Metaxa2, mTAGs, and EukDetect—even for species not represented in the database. Using a similar database of single copy orthologs as EukDetect, we show how a supervised learning approach automates the assessment of which genes and regions are most informative for classification, reducing the need for manual curation, and retaining more sequences in the database. Further, the supervised learning approach of taxaTarget replaces universal classification thresholds (i.e., universally applied to all sequences in a database) with classifiers tuned by sequence region in a data-driven manner. We also describe several novel solutions for effectively handling the scarcity of training data for microeukaryotes which initially led to many classification errors.

4.2 Methods

4.2.1 Definitions

Precision = True Positives / (True Positives + False Positives)

Sensitivity = True Positives / (True Positives + False Negatives)

False Positive Rate = False Positives / (False Negatives + True Negatives)

Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)

4.2.2 Tools selected for benchmarking taxaTarget

To benchmark the performance of taxaTarget we selected several existing tools (Metaxa2, mTAGs, and EukDetect) that were representative of the main marker gene approaches for the taxonomic classification of microeukaryotes in metagenomic data. These approaches can be broadly classified into three categories based upon the marker sequences they use as a database. The first approach uses a database of markers that were originally used as barcodes for amplicon-based metagenomics such as the COI gene and the ribosomal subunits. Here we selected Metaxa2 and the recently published mTAGs. The second approach uses a database of clade-specific marker genes. The only tool we could find that used this approach was MetaPhlan3, which we decided not to include because it had orders of magnitude less eukaryotic species than the other tools (only 122 eukaryotic species as of release mpa_v30). The third approach uses a database of single copy orthologs that are found in most eukaryotes. The only tool that we found using this approach is EukDetect.

4.2.3 Workflow and implementation

The workflow of taxaTarget is shown in Figure 4.1. It is written in Python3 (Python $\geq$ 3.6) and executed via command line. It utilizes a custom database of eukaryotic marker genes, the trained classifiers, and the NCBI taxonomy. It accepts as input raw metagenomic sequencing reads in fasta/fastq format. The reads are first mapped to the database with Kaiju (v1.8.0) [194] to quickly identify reads that might be sequenced from eukaryotic marker genes. The binned reads are then more sensitively aligned with Diamond (v2.0.11) [195] to the marker genes. The Diamond output is then provided as input to taxon-specific classifiers trained on a curated database of eukaryotic sequences. Finally, the results are aggregated and the taxonomic profile of the metagenomic sample is generated.

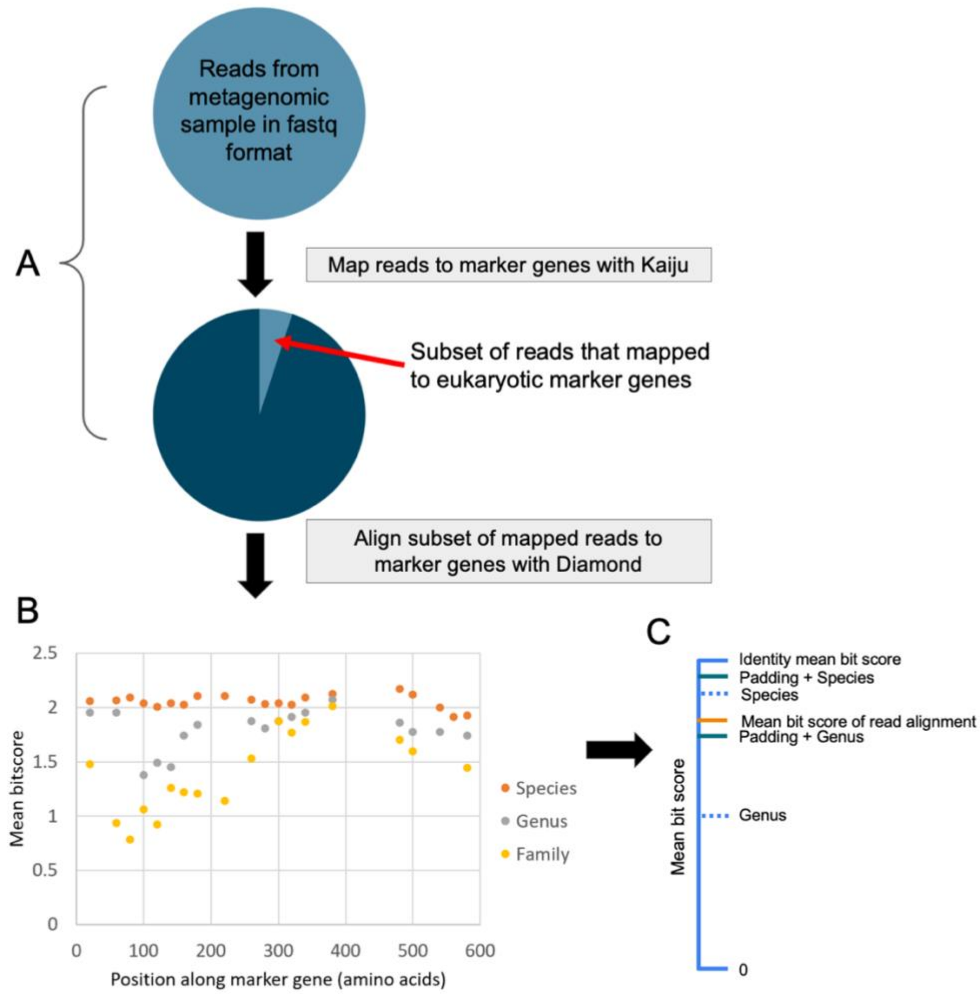

Figure 4.1 Workflow diagram for taxaTarget

A) Metagenomic reads are input to taxaTarget in fastq format. The reads are first mapped to the marker genes with Kaiju—a fast, k-mer approach. The subset of reads that mapped to the marker genes with Kaiju are then more sensitively aligned to the marker genes with Diamond. The mean bit score and start position of the Diamond read alignments are used for classification. B) An example marker gene that a read mapped to. Most windows (every non-overlapping 20 amino acid region) have classification thresholds (Species, Genus, and Family) learned from training, but some regions lacked training data and thus lack thresholds. C) Classification of a read for a window where threshold padding is applied. The start position of the read alignment is used to determine the window used for classification. During training, no negative labelled reads aligned to the window at a rank broader than genus and so the thresholds for Genus and Species get padded (i.e., adjusted upwards). Here, the read gets classified as the genus of the marker gene because the mean bit score of its alignment is above Genus with padding.

4.2.4 Building the marker gene database

The database of taxaTarget was built from protein sequences extracted from the UniRef100 database of UniProt (release-2021_02) [196] to maximize the breadth of species detectable. The UniRef100 database was used, instead of UniProt, because it removes identical sequences and assigns the lowest (most recent) common ancestor to the representative sequences in the database. The ~2.0×10^8 protein sequences in the UniRef100 database were downloaded on April 15, 2021. Approximately 6.4×10^7 eukaryotic proteins were extracted and assigned a taxonomic lineage using the taxonomic identifiers provided by UniRef100 and the NCBI taxonomy. Eukaryotic proteins were excluded if the taxonomic lineage included the keywords "unclassified eukaryotes", "environmental samples", or if the base taxonomic label was at a broader taxonomic rank than species. Eukaryotic marker genes were identified and extracted if they aligned with Diamond to the 255 eukaryotic Benchmarking Universal Single-Copy Orthologs (BUSCO) marker genes (version 5) [188] with an alignment matching or exceeding the BUSCO cutoffs for sequence length and bit score.

4.2.5 Training the taxonomic classifiers

The classifiers of taxaTarget are a set of region-specific functions that use thresholds for family, genus, and species, to classify metagenomic reads. The thresholds for species, genus, and family ($S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$, respectively) are learned for each non-overlapping 20 amino acid window, *W*, of each marker gene, *M*, in the taxaTarget database via training. As an example, the notation for the species threshold of the 140 to 160 amino acid window of marker gene UniRef100_C0H4X5 (its UniRef100 identifier) would be $S_{(UniRef100\_C0H4X5, 140)}$. The thresholds learned during training for each window provide the basis for the taxaTarget classifiers.

To provide an overview, the training process began by decomposing the UniRef100 database into a set of reads. Each read was aligned to the taxaTarget database, yielding one or more alignment scores depending on how many marker genes it aligned to. Per alignment, each read was assigned a label, either “species”, “genus”, ”family”, or “negative”, depending on its taxonomic relationship to the marker gene. A read that aligned to a marker gene from a broader rank than family was considered a negative. The set of labelled reads that mapped within the window of a marker gene were used to determine the thresholds for that window.

The alignment score used for training and classification was the mean bit score—the bit score of a read alignment divided by the alignment length. The mean bit score was used because there is an approximately linear relationship between the alignment bit score and alignment length [192], i.e., reads of different lengths, sequenced from the same gene, will align back to the marker gene with different bit scores, but similar mean bit scores. Therefore, training on the mean bit score allows for standardized classification of reads of different lengths.

In detail, training began by aligning all UniRef100 reads (reads are all 70 amino acid long k-mers, with step size 10, extracted from UniRef100), to the marker genes using Diamond BLASTP. The procedure for identifying the species, genus, and family thresholds ($S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$, respectively) for each window, W, of each marker gene, M, in the taxaTarget database is as follows. To provide a specific example for ease of understanding, let us say we are trying to learn $S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$ for the 140 to 160 amino acid window of marker gene UniRef100_C0H4X5, i.e., $S_{(UniRef100\_C0H4X5, 140)}$, $G_{(UniRef100\_C0H4X5, 140)}$, $F_{(UniRef100\_C0H4X5, 140)}$.

1) If a read alignment to UniRef100_C0H4X5 started or ended within the 140 to 160 amino acid window (extending towards the end of the gene) with a minimum bit score of 60 and alignment length of at least 30 amino acids,
    a) The mean bit score of the read alignment was extracted
    b) A taxonomic label ("species", "genus", "family", or "negative") was assigned depending on the taxonomic relationship of the read to the marker gene (*Plasmodium falciparum* for UniRef100_C0H4X5)
2) $S_{(UniRef100\_C0H4X5,\ 140)}$, $G_{(UniRef100\_C0H4X5,\ 140)}$, $F_{(UniRef100\_C0H4X5,\ 140)}$ were set to the minimum mean bit score of the reads mapped and corresponding to that taxonomic rank
3) Thresholds were adjusted or removed based on the following conditions
    a) A threshold was removed if there was no training data corresponding to its taxonomic rank
    b) If the mean bit score of a broader taxonomic rank was higher than the threshold of a more specific taxonomic rank, that threshold was either adjusted to that mean bit score or removed if the mean bit score was higher than the threshold of the next most specific rank
    c) The window was only kept in taxaTarget for classification if,
        i) $S_{(UniRef100\_C0H4X5,\ 140)}$ had a value (in some cases reads did not map to a window)
        ii) $S_{(UniRef100\_C0H4X5,\ 140)} > \{G_{(UniRef100\_C0H4X5,\ 140)}, F_{(UniRef100\_C0H4X5,\ 140)}, \max(\text{negatives})\}$
4) The process was repeated for every window of every marker gene in the database.

At the end of the training process, the values for $S_{(M,W)}$, $G_{(M,W)}$, and $F_{(M,W)}$ were defined for each window, for each marker gene, subject to the filtering parameters described above. This provided the basis for the classifiers used by taxaTarget for classification.

4.2.6 Taxonomic classification of metagenomic reads

4.2.6.1 Classifying individual reads

Taxonomic classification with taxaTarget begins by aligning metagenomic reads to the taxaTarget database with Diamond BLASTX, yielding alignment scores. The highest mean bit score alignment(s) of individual reads are used for classification. If a read aligns with 100% identity to multiple marker genes from different species, the read is classified at the lowest common ancestor of the respective hits. Otherwise, individual reads are classified as follows:

1. The alignment of a read to a marker gene must be at least 30 amino acids long (gaps are counted)
2. A read alignment must either start or end in the window of a classifier and extend toward the end of the marker gene. If a read alignment overlaps multiple windows, the window used for classification is, window = floor(min(start, end) / 20). If a classifier does not exist for that window, the read is not classified.
3. The classifier for a window consists of a function that uses the learned classification thresholds for species, genus, and family, $S_{(M,W)}$, $G_{(M,W)}$, and $F_{(M,W)}$, respectively, where M is the specific marker gene the read aligned to and W is the window.
4. The mean bit score of the read alignment is provided as input to the classifier
5. Classification
    a. If the mean bit score $\geq S_{(M,W)}$, classify the read at the species level
    b. If the mean bit score $\geq G_{(M,W)}$, classify the read at the genus level

c. If the mean bit score ≥ $F_{(M, W)}$, classify the read at the family level

d. Else, the read is not classified

4.2.6.2 Generating the taxonomic profile of the metagenomic sample

Once the individual reads have been taxonomically classified, taxaTarget uses several functions to analyze the pattern of mapped reads for each taxon and to filter out false positives from the taxonomic profile. The sources of false positives were identified through experiments or reported in the literature previously [184]. Each parameter has a default setting that can also be set by the user.

1. Thresholds determined from training data that lacked examples from broader taxonomic ranks or negatives are frequently overly permissive leading to false positive classifications. To address this factor, taxaTarget relies on "padding," a user tunable parameter, to adjust the thresholds higher (requiring a higher mean bit score to classify a sequence). The default padding added (0.5 by default) was determined via parameter sweep analyses detailed later in the Methods. The following padding strategy is used:
    a. By default, $T_{padding} = T + 0.5(I - T)$, where T is the threshold (options are $S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$) before padding, I is the mean bit score of the identity alignment for the window, and $T_{padding}$ is the threshold value after padding is added
    b. Padding is added using the following series of rules. Once a rule is met, the subsequent rules are not applied. During training, if no reads aligned to a marker gene,
        i. At a rank broader than family, then padding is added to $S_{(M, W)}$, $G_{(M, W)}$, and $F_{(M, W)}$

ii. At a rank broader than genus, then padding is added to $S_{(M, W)}$ and $G_{(M, W)}$

iii. At a rank broader than species, then padding is added to $S_{(M, W)}$

2. As reported in [184] and supported by our observations, false positives frequently entail just one or two reads mapping to a single marker gene of an organism. To minimize such errors, taxaTarget requires that at least 3 reads be classified per taxonomic unit, and that reads map to at least 3 marker genes within a taxonomic unit for it to be reported in the output.

3. If sequencing samples the genomes in a metagenome randomly, then the reads should approximately cover a genome uniformly. False positives are often characterized by many reads mapping to a small subset of the marker genes from a particular taxon, perhaps due to a conserved domain. taxaTarget relies on outlier detection to filter out marker genes that have more mapped reads than expected for each taxon. The cutoff for the maximum expected number of mapped reads for a marker gene is approximated as follows:

    a. Let $X = \{x_1, x_2, ..., x_N\}$, where X is a vector of counts and x_i is a single draw, sampled with replacement, from the set of marker genes, $\{m_1, m_2, ..., m_K\}$, where N is the total number of reads mapped to the marker genes, and K is the total number of marker genes for a taxon. $C = \text{mode}(X) \times 2$.

    b. Any marker gene with more mapped reads than the mean of C after 100 experiments is excluded for the taxon being considered.

4. A source of error, also identified in [184], is that more species of a genus get detected than are present in the sample. Sometimes, this is due to the species lacking a complete set of marker genes in the taxaTarget database. To address such situations, taxaTarget

examines the pattern of read mappings to the marker genes from the equivalent species using an extension of the strategy reported in [184]. First, a primary species is defined as the species with the most mapped reads that passes all previous filters. The primary species is assumed present, and the other species are compared to it to determine if they are present as well or false positives. For demonstration purposes, let us consider two scenarios (that taxaTarget handles) where two species, A and B, within the same genus have mapped reads, where A is the primary species.

a. If the reads of species A and B map to non-overlapping sets of marker genes, the read counts are summed and the taxonomic profile reports both species as a single result indicating there is not enough information to determine whether either or both are present.
b. If the reads of species A and B map to partially overlapping sets of marker genes, species B is considered a false positive if the ratio of reads mapped (normalized by the number of marker genes) to shared genes versus not-shared genes is less than 1, with a tolerance of 10%.

4.2.7 Simulated data sets

The simulated datasets were created using ART (Mount Rainier version) [197] (using default parameters except for, Illumina HiSeq 2500, single-end, 150 bp, 1X depth of coverage) for the genomes described in Table 4.1.

Table 4.1 The composition of the simulated metagenomic datasets

Dataset/Analysis	Phylogroup	Number of genera	Number of species/assemblies	Number of reads simulated

non-microeukaryote	Animals	383	383	10,022,536
	Archaea	139	139	2,511,659
	Bacteria	2,761	2,761	10,653,348
	Plants	77	77	10,140,188
	Viruses	2,171	2,171	464,121
microeukaryote_in_DB	Protists, Fungi (species represented in all tool databases)	39	54	7,536,554
microeuryote_not_in_DB	Protists, Fungi (species not represented in database of any tool)	83	83	32,052,306
Limit of detection analysis	*Plasmodium falciparum*	1	1	100,000
	Phytophthora sojae	1	1	100,000
	Aspergillus niger	1	1	100,000
	Blastocystis subtype 1	1	1	100,000

For the *non-microeukaryote* dataset, the GenBank assembly report (available at https://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/assembly_summary_genbank.txt) was parsed with a custom Python script to download one assembly per genus for each of the following phylogroups: *metazoa* (animals), *embryophyta* (plants), archaea, bacteria, and viruses. Per genus, the assemblies were first prioritized by assembly level, i.e., complete genome, chromosome, scaffold, contig. Then an assembly was randomly selected from the highest assembly level available for that genus. For analysis, simulated reads were combined into metagenomes by phylogroup and randomly down sampled to approximately 10 million reads. All reads simulated from the archaea and virus assemblies were used as they were already fewer than 10 million.

Assemblies from 54 species for the *microeukaryote_in_DB* dataset were downloaded from the NCBI assembly database after searching for “eukaryota” and filtering for “Latest RefSeq”, “Protists”, “Fungi”, “Complete genome”, “Chromosome”, “Representative.” Here, we only used high quality assemblies to reduce the likelihood of contaminating sequences in the assemblies which could affect the classification results. The assemblies were further filtered to only include species that had sequences in the databases of each tool tested (taxaTarget, EukDetect, mTAGs, and Metaxa2). The read sets were analyzed individually as well as a metagenome mixture where all the reads had been combined into a single fastq file.

The assemblies for the 83 species in the *microeukaryote_not_in_DB* were downloaded using the GenBank assembly report and a custom Python script that found species without sequences in any of the databases used by mTAGs, Metaxa2, EukDetect, and taxaTarget, but that did have sequences from the same genus in these databases. Each read set was analyzed individually.

4.2.8 Analyses using simulated data

4.2.8.1 Parameter sweep analysis for adjusting classification thresholds

As mentioned earlier, thresholds learned from training data that lacked examples from broader taxonomic ranks frequently allowed false positive classifications. To address this, taxaTarget relies on “padding” these thresholds, a user tunable parameter, to adjust the thresholds higher. The default amount of padding to add to thresholds was assessed via a parameter sweep analysis. The values tested for padding included (0.0, 0.1, 0.2, …, 1.0). The experiment involved three simulated datasets which were useful for assessing different performance metrics of taxaTarget.

First, the *non-microeukaryote* dataset was used to assess the false positive rate of taxaTarget—both with and without applying outlier detection (see above). The false positive rate was calculated as the proportion of reads from non-protist/non-fungi phylogroups (i.e., animals, archaea, bacteria, plants, and viruses) that were misclassified as a protist or fungus.

Second, the metagenome of the *microeukaryote_in_DB* dataset was used to assess the relationship between the precision and the sensitivity of taxaTarget for species with sequences in the database. The precision was calculated using the number of true positive and false positive reads classified at the species level. The sensitivity of taxaTarget was calculated using the number of true positive and false negative species detected in the metagenome.

Third, the *microeukaryote_not_in_DB* dataset was used to assess the relationship between the precision and sensitivity of taxaTarget for species with no sequences in the database, but that have relatives from the same genus within the database. The precision and sensitivity were calculated as before but at the genus level.

4.2.8.2 Limit of detection analysis

Simulated read sets, meant to mimic metagenomic samples where protist and fungal species occur at low abundance, were created as a control to test the limit of detection of the tools. The reads were simulated from the GenBank assemblies of three protists—*Blastocystis subtype 1* (GCA_001651215.1), *Plasmodium falciparum* (GCA_000002765.2), and *Phytophthora sojae* (GCA_000149755.2)—and one fungus—*Aspergillus niger* (GCA_000002855.2)—representing a range genome sizes (1.6×10^7 bp to 8.3×10^7 bp). Six read sets were created from each assembly, composed of 1, 10, 100, $1x10^3$,$1x10^4$, and $1x10^5$ reads using a custom Python script to randomly select from the ART simulated reads. Every tool tested had reference sequences from these species in their database. Each read set was classified separately with the tools and the number of reads correctly classified at the genus level were recorded.

4.2.8.3 Testing the false positive rate of the tools for non-protist/non-fungi species

The *non-microeukaryote* dataset (Table 4.1) was used to test the false positive rate of the tools. The false positive rate was calculated as the proportion of reads from non-protist/non-fungi phylogroups (i.e., animals, archaea, bacteria, plants, and viruses) that were misclassified as a protist or fungi. The mean false positive rate for all phylogroups was normalized by 10 million to derive the mean number of false positive reads per 10 million reads classified.

4.2.8.4 Measuring the precision and sensitivity of the tools for protist and fungal species

The *microeukaryote_in_DB* dataset (Table 4.1) was used to measure the precision of the tools when classifying protist and fungal species with reference sequences in the database. The tools were first applied to each read set individually to provide a baseline for the precision of the tools and then to the combined read set to assess the precision of the taxonomic profile when multiple species were present (sometimes multiple species from the same genus). The precision was calculated separately at the taxonomic levels of family, genus, and species. The sensitivity of the tools was defined at the family, genus, and species levels as the proportion of the total number species in the dataset where at least one read was classified correctly at the respective taxonomic levels.

The *microeukaryote_not_in_DB* dataset (Table 4.1) was used to measure the precision of the tools when classifying protist and fungal species with no reference sequences (although there were reference sequences from other species within the corresponding genera) for the species in the databases used by the tools. The precision and sensitivity of the tools were measured as mentioned in the previous paragraph except only at the taxonomic levels of family and genus as the respective species are not represented in the reference collection.

4.2.9 Analysis of real metagenomic data

The accuracy of the tools for detecting known protist species in a microbiome was quantified using 57 Cameroonian shotgun metagenomic gut samples [91]. Polymerase chain reaction (PCR) had previously been performed on the samples to detect Blastocystis and microscopy to detect Entamoeba. Additionally, the previous study had identified several target protist species by mapping the metagenomic reads to a database of their genomes or 18S rRNA sequences [91]. We summarized and compared the taxonomic profiles of each tool with the findings of the original study. For *Blastocystis*, the PCR results were used as a "gold standard" for measuring the accuracy of the tools for detecting this parasite.

4.3 Results

4.3.1 The taxaTarget database

The reference database of taxaTarget is composed of 877,724 UniRef100 proteins which were identified as orthologs of the 255 BUSCO eukaryotic marker genes. The database sequences originate from 1,319 families, 2,564 genera and 4,232 species. As seen in Figure 4.2B, many species had few marker genes (e.g., 1,152 species only had one marker gene identified); however, 2,619 species had over 50 marker genes and the median number of marker genes per species was 152.

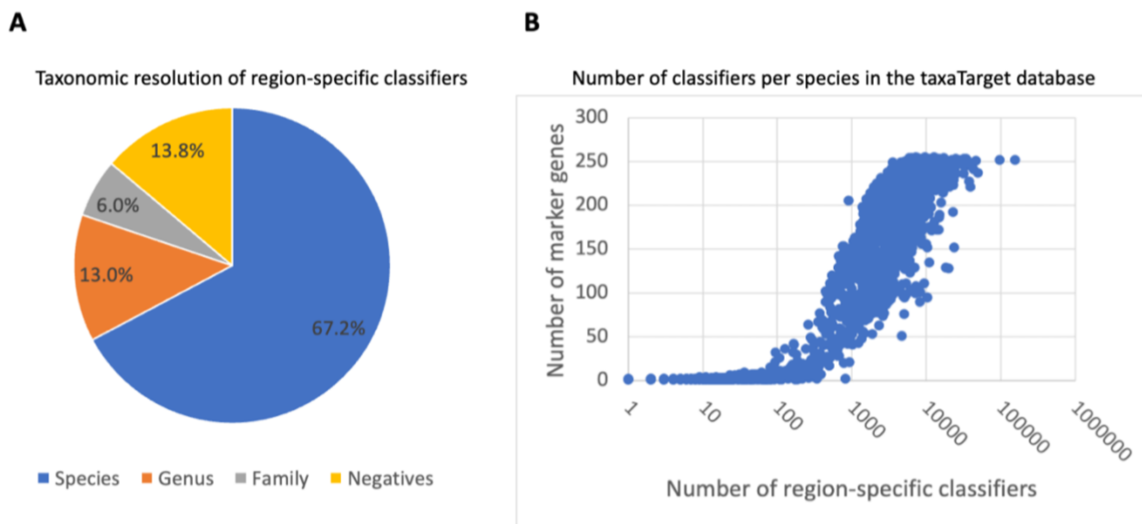

Figure 4.2 The region-specific taxonomic classifiers of taxaTarget
A) The pie chart shows the maximum taxonomic resolution of the regions learned by training. Nearly 1/3 of the 22.7x107 non-overlapping 20 amino acid regions (from 8.8x105 proteins) were not useful for species level classification because the highest bit score alignments were from broader taxonomic ranks. B) The scatter plot shows the number of BUSCO marker genes (255 is a complete set) versus the number of region-specific classifiers for all 4,232 species in the taxaTarget database.

4.3.2 Statistics about the trained classifiers of taxaTarget

Classifiers were trained for all non-overlapping 20 amino acid regions (22.7×10^7 potential regions) per protein in the database. About ⅓ of the regions (7.4×10^6) were excluded after training because the highest mean bit score alignments for family, genus, and species did not follow the pattern of species > genus > family (Figure 4.2A). For the remaining ~⅔ of regions, training data was often sparse i.e., 22.7%, 64.2%, and 68.1% of the regions lacked negative-labeled, family-labeled, or genus-labeled data, respectively. The number of region-specific classifiers varied by orders of magnitude between marker genes, from 2.8×10^3 to 1.5×10^6 with a median of 4.3×10^4. There was also great variability per species, from 1 (e.g., *Phytophthora andina*) to 1.6×10^5 (e.g., *Aureobasidium pullulans*) with a median of 2.1×10^3 region-specific classifiers per species. Furthermore, there were cases where the mean bit score showed high variability across windows of a marker gene within a single rank; the maximum mean bit score range observed was 1.94 across the windows of the Brix domain-containing protein of *Hortaea werneckii* (A0A3M7GIU8).

4.3.3 Parameter sweep analyses: Padding thresholds trained with sparse data

The first parameter sweep analysis assessed how threshold padding, with and without outlier detection, affected the number of non-microeukaryote reads falsely classified as microeukaryotic (protist or fungal) species (Figure 4.3A) by taxaTarget. Outlier detection refers to whether taxaTarget is set to remove marker genes with more mapped reads than expected (see Methods).

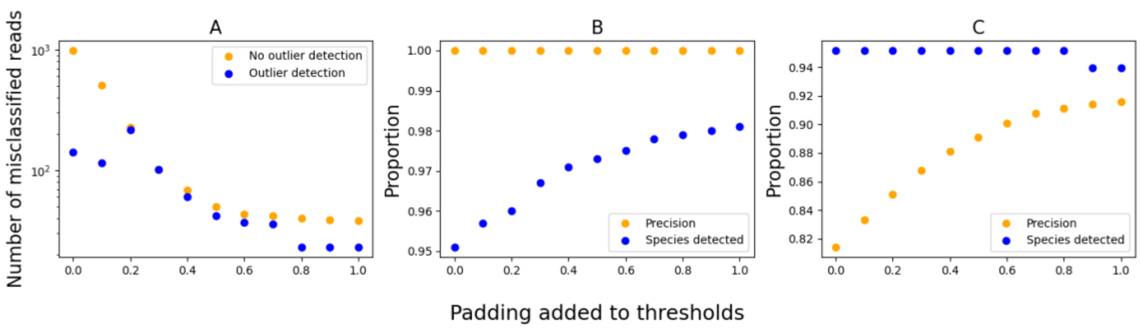

Figure 4.3 Results of parameter sweep analysis for taxaTarget
A) The *non-microeukaryote* metagenome was used to assess how the padding of thresholds affected the number of non-microeukaryote reads falsely classified as microeukaryotic (protist or fungal) species. Outlier detection refers to whether taxaTarget is set to remove marker genes with more reads mapped than expected—if sequencing randomly samples the genomes in a metagenome and that read coverage should approximate a uniform distribution. B) The *microeukaryote_in_DB* metagenome was used to assess the effect of threshold padding on the precision and the number of detected species (n = 54) at the species level. C) The *microeukaryote_not_in_DB* dataset was used to assess the effect of threshold padding on the precision and the number of detected species (n = 83) at the genus level.

Irrespective of padding, taxaTarget always performed better when using outlier detection, especially at low threshold padding, i.e., 0.2 or less (Figure 4.3A). Without outlier detection, the number of false positive reads per 10 million reads classified decreased from 179 to 7 as the threshold padding increased from 0 to 1. With outlier detection, the same value over the same range decreased from 26 to 4. From 0.1 to 0.2 padding there was an unexpected increase in the number of false positive reads classified as *embryophyta*. This was caused by several species where the pattern of mapped reads to marker genes was not uniformly distributed at 0.1 padding—resulting in them being filtered out of the taxonomic profile by taxaTarget. However, at 0.2 padding the number of reads mapped to marker genes was greatly reduced and approximated uniform coverage—resulting in them being wrongly included in the taxonomic profile. The phylogroups that contributed to the most false positives, irrespective of padding level, were the bacteria and embryophyta. For bacteria, misclassified reads were most frequently classified as *fungi*, *chlorophyta*, and *choanoflagellata*. For example, at 0.5 padding there were 15 misclassified reads, of which 7, 5, and 4 were assigned to *fungi*, *chlorophyta*, and *euglenozoa*, respectively. For *embryophyta*, misclassified reads were most frequently classified as *chlorophyta* and *charophyceae*—green algae. For example, at 0.5 padding there were 23 misclassified reads, 20 were assigned to *chlorophyta* (the other 3 were assigned to *stramenopiles*). The viruses, archaea, and *metazoa* did not produce false positives, only did so at a low rate (irrespective of threshold padding), or only at a padding of 0.1 or lower, respectively.

The second analysis used the *microeukaryote_in_DB* metagenome to assess the effect of padding thresholds on the precision and sensitivity of taxaTarget at the species level for species with sequences in the reference database (Figure 4.3B). The sensitivity was 1.0 regardless of padding, i.e., all species were detected at the species level in the metagenome. With respect to precision, as padding increased from 0 to 1 there was a slight increase from 0.951 to 0.981, respectively.

The third analysis used the *microeukaryote_not_in_DB* dataset to assess the effect of padding thresholds on the precision and sensitivity of taxaTarget for species without sequences in the database (Figure 4.3C). The precision was assessed at the genus level. Similarly, the sensitivity was calculated as the proportion of species detected at the genus level. Overall, the sensitivity was 0.952 from 0 to 0.8 padding and 0.94 thereafter. The precision continuously increased from 0.814 to 0.916 from a padding of 0 to 1, respectively.

Together, the parameter sweep analyses led us to choose a threshold padding of 0.5 as the default setting for taxaTarget, but users can adjust this parameter.

4.3.4 Limit of detection analysis

The limit of detection for each tool was tested by classifying read sets simulated from genomes, where the number of reads varied by 6 orders of magnitude (1, 10, 100, $1x10^3$, $1x10^4$, $1x10^5$). The datasets were simulated from the genomes of two protist species (*Plasmodium falciparum* and *Phytophthora sojae*) and one fungal species (*Aspergillus niger*). The genomes ranged in size from 23, 34, and 83 Mbp for *Plasmodium falciparum*, *Aspergillus niger* and *Phytophthora sojae*, respectively. Each species had reference sequences in the database of each tool. The sensitivity of a tool was measured as the number of reads correctly classified at the genus level.

Amongst the tools, taxaTarget showed the greatest sensitivity (Figure 4.4), only requiring 100 reads to detect *Aspergillus* and 1,000 reads to detect *Phytophthora*, *Blastocystis*, and *Plasmodium*. EukDetect required 10,000 reads to detect *Aspergillus, Blastocystis*, and *Phytophthora* and 100,000 reads to detect *Plasmodium*. Metaxa2 and mTAGs showed the least sensitivity, unable to detect *Phytophthora* given 100,000 reads, and requiring 10,000 to 100,000 reads to detect *Aspergillus*, *Blastocystis*, and *Plasmodium*.

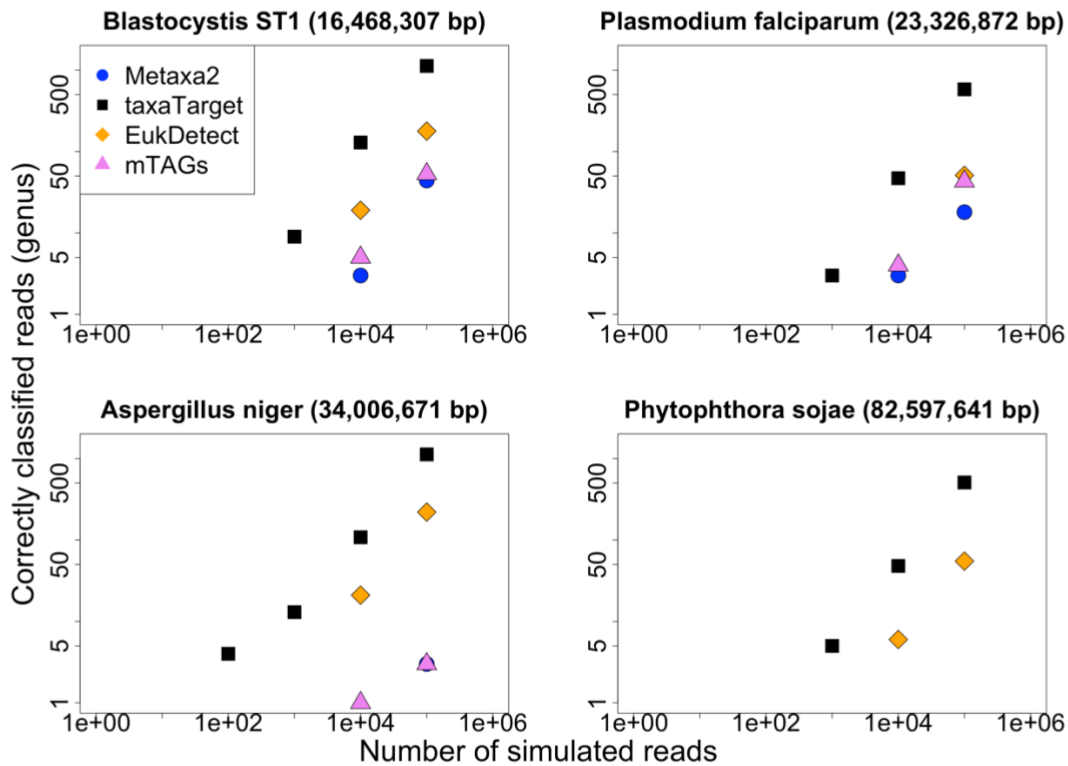

Figure 4.4 Limit of detection analysis

The limit of detection for each tool was tested using simulated data where the number of reads varied by 6 orders of magnitude (1, 10, 100, $1x10^3$, $1x10^4$, $1x10^5$). A read was considered correctly classified if the tool assigned it to the correct genus. The reads were simulated from 4 genomes which varied in length from ~16.6 Mbp to ~82.6 Mbp. The genomes came from 3 protist species (*Blastocystis subtype 1*, *Plasmodium falciparum* and *Phytophthora sojae*) and 1 fungal species (*Aspergillus niger*)—genome size in parentheses besides the species name.

4.3.5 False positive rate when classifying simulated metagenomes composed of non-fungi/non-protist phylogroups

Microbiomes can be composed of hundreds or thousands of microbial species—usually only a small proportion are microeukaryotes (mostly protists and fungi) [179, 198]. The non-microeukaryote dataset was used to quantify the false positive rate of the tools when classifying non-protist/non-fungal phylogroups.

As seen in Table 4.2, EukDetect had the lowest overall false positive rate of zero. The mean number of false positive reads per 10 million reads classified for Metaxa2, taxaTarget, and mTAGs was 4, 8, and 15 respectively. For taxaTarget, the false positive rate was mostly driven by bacteria classified as *Choanoflagellata* and plants classified as green algae (*Charophyceae* and *Chlorellales*).

Table 4.2 The false positive rate of the tools for the simulated metagenomes from non-protist/non-fungi phylogroups

Tool	Archaea	Bacteria	Plants	Animals	Viruses	Mean number of false positive reads per 10 million reads classified
EukDetect	0	0	0	0	0	0
Metaxa2	0	6.6×10^{-7}	5.9×10^{-7}	2.0×10^{-7}	0	4
taxaTarget	1.2×10^{-6}	1.5×10^{-6}	2.3×10^{-6}	0	0	8
mTAGs	4.0×10^{-7}	3.7×10^{-6}	8.9×10^{-7}	2.0×10^{-7}	0	15

4.3.6 Classification performance for "known" fungal and protist species

The precision and sensitivity of the tools for classifying protist/fungal species, with representative sequences in their databases, was tested under two conditions using the *microeukaryote_in_DB* dataset. First, the classification performance of the tools was tested on simulated genomic read sets from 54 species to provide a baseline. Second, a 54 species metagenome (combining the genomic read sets into a single sample) was used to measure the precision of the tools on a complex, multispecies sample where 24 of the 54 species came from a genus with more than one species in the metagenome. We note that the results of mTAGs could not be compared with the other tools at the species level because mTAGs uses operational taxonomic units (OTU). OTUs are taxonomic groups defined by the clustering of marker sequences, which can evolve at different rates, and thus can have variable taxonomic resolution e.g., strain, species, subgenus.

When classifying genomic read sets (Table 4.3) all tools showed a precision of at least 95%, although taxaTarget often had the highest. However, only taxaTarget had 100% sensitivity i.e., correctly identified all species. EukDetect, had the second highest sensitivity (90.7%) only misclassifying 5 species as other species. In contrast, mTAGs and Metaxa2 had the lowest sensitivity, classifying orders of magnitude less reads and identifying the least genera. For example, Metaxa2 only identified 31.5% of the species and only classified 320 reads (out of a total of 7.5 million) at the species level.

Table 4.3 Precision of tools when classifying species represented in their databases.

Precision of tools when classifying read sets (1X depth of coverage) simulated from the genomes of 54 protist/fungal species with reference sequences in the database of each tool. In the experiment column, genome indicates the read sets were analyzed individually, whereas metagenome indicates the read sets were combined into a single sample that was analyzed.

*Metaxa2 uses a custom and older taxonomic database which often does align with the current NCBI taxonomy at the family level. If the family label did not align with the current NCBI taxonomy it was counted as incorrect.

**Below genus, mTAGs classifies reads to OTUs and does not provide species level classifications

Experiment	Tool	Precision at family level (%)	Precision at genus level (%)	Precision at species level (%)	Percentage of species detected (n = 54)	Reads classified at species level (n = 7.5×10^6)
Genome	taxaTarget	99.0	98.9	97.3	100.0	35,950
	EukDetect	96.0	96.0	95.3	90.7	51,489
	Metaxa2	66.8*	97.9	97.4	31.5	320

	mTAGs	98.8	96.3	NA	NA	703 (OTU**)
Metagenome	taxaTarget	99.0	99.0	97.3	100.0	36,182
	EukDetect	96.2	96.2	95.0	68.5	44,823
	Metaxa2	76.3*	99.2	97.0	31.5	297
	mTAGs	97.5	94.9	NA	NA	713 (OTU**)

When classifying the combined metagenome (Table 4.3), all tools still showed high precision (≥ 95%). Metaxa2 and taxaTarget had the highest precision at the genus (~99%) and species (~97%) levels, but taxaTarget was still the only tool with 100% sensitivity. EukDetect showed a substantial loss in sensitivity (only detecting one species per genus when multiple species were present), detecting 68.5% of the species, in contrast to 90% when data from each species was analyzed separately. Like the previous experiment, mTAGs and Metaxa2 still showed the lowest sensitivity.

4.3.7 Classification performance for "unknown" fungal and protist species

Next, we tested if the tools could detect protist/fungal species that were not represented in their databases, but that had marker sequences from the same genus. We tested the tools using the *microeukaryote_not_in_DB* dataset which consisted of simulated genomic read sets from 83 species. The 83 species had no reference sequences in the databases of any tool but had sequences from species related at the genus level (Table 4.4). We again note that the results of mTAGs could not be compared with the other tools at the species level (refer to the first paragraph of the previous section for a detailed explanation).

Table 4.4 Analysis of 83 single species simulated read sets.
The 83 protist/fungal species had no reference sequences in the database of the tools, but that had sequences from species related at the genus level.

Tool	Species detected at family level (n = 83)	Species detected at genus level (n = 83)	Sensitivity family (%)	Sensitivity genus (%)	Precision family (%)	Precision genus (%)
taxaTarget	81	79	97.6	95.2	94.5	89.1
EukDetect	44	42	51.8	49.4	99.9	90.4
Metaxa2	23	14	27.7	16.9	90.3	41.6
mTAGs	39	30	47.0	36.1	88.1	50.8

In terms of sensitivity, taxaTarget identified the correct family and genus for over 95% of the species. In contrast, all the other tools had a sensitivity of ~50% or less. EukDetect was the most precise tool, assigning the correct family and genus to 99.9% and 90.4%, respectively, of the reads it classified. taxaTarget showed the second highest precision, assigning 94.5% and 89.1% of the classified reads to the correct family and genus, respectively. In contrast, Metaxa2 and mTAGs had ~90% precision at the family level and ~50% at the genus level.

4.3.8 Detection of known genera in the Cameroonian gut metagenomes

The accuracy of the tools for detecting known protist species in a microbiome was quantified using 57 Cameroonian shotgun metagenomic gut samples (Table 4.5). PCR had previously been performed on the samples to detect *Blastocystis* and microscopy to detect *Entamoeba*. Additionally, several target protist species had been detected by mapping the metagenomic reads to a database of their genomes or to their 18S rRNA [91].

Table 4.5 Detection of *Blastocystis* in 57 Cameroonian gut samples.
PCR(+) indicates that PCR detected *Blastocystis* in the sample. Tool(+) indicates the tool detected *Blastocystis* in the metagenomic sample.

Tool	PCR(+) Tool(+)	PCR(-) Tool(-)	PCR(+) Tool(-)	PCR(-) Tool(+)	Accuracy (%)
Metaxa2	48	7	1	1	96.5
mTAGs	48	6	1	2	94.7
taxaTarget	44	7	5	1	89.5
EukDetect	40	7	9	1	82.5

For *Blastocystis*, the PCR results were used as the "gold standard" for the analysis. The accuracy of the tools was assessed at the genus level because the databases of the tools did not contain all the *Blastocystis* species detected by PCR. Metaxa2 had the highest agreement with the PCR results, 96.5%, followed by mTAGs (94.7%), taxaTarget (89.5%) and EukDetect (82.5%). All the tools detected *Blastocystis* in one sample that was PCR negative. For the 5 PCR positive samples where taxaTarget did not detect *Blastocystis*, three samples only had one or two classified reads—taxa with fewer than 3 classified reads are removed from the taxonomic profile by taxaTarget. The other two samples had three and seven reads, respectively, but they mapped to fewer than three marker genes, the minimum number required by taxaTarget to be included in the taxonomic profile.

For *Entamoeba*, microscopy had previously been found less sensitive than mapping the metagenomic reads to reference genomes—33% and 63% of samples were found positive using microscopy and read mapping, respectively [91]. For our analysis, *Entamoeba* was identified in 54% (mTAGs), 37% (Metaxa2), 23% (taxaTarget), and 0% (EukDetect) of the samples, respectively. All *Entamoeba* positive samples were a subset of those identified with metagenomic read mapping in the previous study. However, *Entamoeba* was only identified in 84% (mTAGs), 69% (taxaTarget), and 67% (Metaxa2) of the *Entamoeba* positive samples determined via microscopy. Many of the samples where taxaTarget did not detect *Entamoeba* had been filtered from the taxonomic profile because too few reads mapped, or the reads mapped to too few marker genes. Before filtering, taxaTarget had identified *Entamoeba* in 95% of microscopy positive and 89% of read mapping positive samples.

Compared to the previous analysis [91], none of the tools detected *Endolimax* or *Enteromonas* at the genus level. However, despite not having database sequences from *Enteromonas*, taxaTarget still detected *Hexamitidae* (the family of *Enteromonas*) in 3 of 4 samples where *Enteromonas* had been identified previously. Further, the previous study had also identified *Giardia* in two samples—mTAGs detected *Giardia* in both whereas taxaTarget either detected *Giardia* or *Hexamitidae* (the family of *Giardia*).

4.3.9 Comparison of the taxonomic profiles for the 57 Cameroon gut samples

The total diversity of microeukaryotes identified in the Cameroon human gut metagenomes was compared for the four tools (Figure 4.5). Overall, 59 genera were detected—EukDetect, taxaTarget, Metaxa2, and mTAGs detected 21%, 35%, 39%, and 46% of the genera, respectively. Amongst these genera, only *Blastocystis* was detected by all tools—also had the highest occurrence in the samples. However, taxaTarget had the greatest overlap with the other tools, sharing 9 genera with EukDetect, 6 with mTAGs, and 3 with Metaxa2. The 3 genera taxaTarget did not share with EukDetect were either not in the database of taxaTarget (*Kodamaea*) or were not reported in the results for having too few marker genes in the database (*Volvariella*) or too few mapped reads (*Bactrocera*). For the 18 genera not shared with Metaxa2, 10 were not in the database of taxaTarget and 6 were not reported in the results for having too few marker genes in the database, too few mapped reads, or too few marker genes with mapped reads. Further, although *Aplanochytrium* was not detected at the genus level, its family was detected. Only the genus *Amphora* showed no signs of detection at the family or genus levels despite having sufficient marker genes in the taxaTarget database to be reported in the taxonomic report. For the remaining 18 genera not shared with mTAGs, mTAGs had detected each in a single sample by a single read. Six of these genera were not in the taxaTarget database, five were in the database but were not detected, two were detected at the family level, and five were not reported because too few reads mapped.

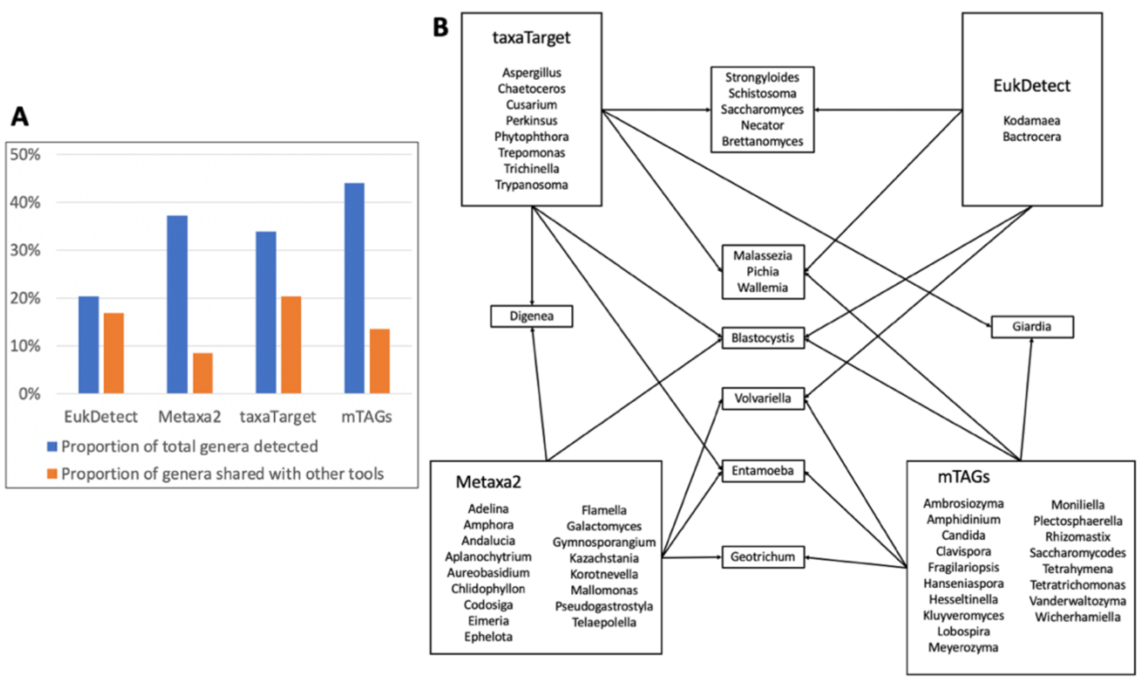

Figure 4.5 Genera of microeukaryotes found by the tools in the 57 Cameroonian gut samples.
A) In total, 59 genera were detected by the tools. None of the tools detected all the genera, but mTAGs detected 44%. B) taxaTarget had the greatest overlap with other tools for detected genera. Only 24% of the genera were detected by more than one tool—taxaTarget detected 86% of them. taxaTarget and EukDetect, which use a similar database of marker genes, shared 9 detected genera, the most between any tools. Metaxa2 and mTAGs, despite using a similar database of marker genes, only had 4 detected genera in common.

Consistent with other studies of the human gut eukaryome [198], no genus was found in all samples and most genera were rare. For perspective, there were only three (EukDetect), four (mTAGs), and five (Metaxa2 and taxaTarget) genera identified in five or more samples. For these genera, taxaTarget contained all detected by EukDetect and all but one by Metaxa2, *Korotnevella*, and by mTAGs, *Hesseltinella*.

Of the 20 genera detected by taxaTarget in the 57 Cameroonian gut samples, 16 had been observed in the human gut microbiome before (Table 4.6). Of the remaining four, *Chaetoceros*, *Perkinsus*, and *Wallemia* potentially could have been transiently introduced to the gut microbiome if a person consumed foods that carried them. Amongst the genera identified by taxaTarget were parasites with clinical relevance to humans such as *Trichinella*, *Schistosoma*, *Strongyloides*, *Digenea*, *Necator*, *Trypanosoma*, and *Giardia*.

Table 4.6 The 20 genera detected by taxaTarget in the 57 Cameroonian gut microbiome samples

Cameroon (prevalence in 57 samples)	Observed in human gut before	Relationships to humans
Blastocystis (77%)	yes	commensal/parasite protist that can consume bacteria in gut [199]
Entamoeba (23%)	yes	commensal/parasitic amoeba that can consume bacteria in gut [200]
Malassezia (19%)	yes	broadly dispersed fungi, once thought to be restricted skin [201]
Saccharomyces (19%)	yes	broadly distributed yeast, some species are commonly found in food [202]
Chaetoceros (12%)	no	none, primarily marine diatoms; might transiently occur in human gut if a colonized shellfish is eaten [203]
Pichia (7%)	yes	yeast commonly found in food [204]

Strongyloides (5%)	yes	parasitic nematode [205]
Trypanosoma (4%)	yes	protist parasite [206]
Phytophthora (4%)	yes	plant pathogen [207]
Aspergillus (4%)	yes	diverse genus of mold consisting of several hundred species [202]
Digenea (4%)	yes	parasitic flatworm (fluke) [208]
Schistosoma (4%)	yes	parasitic flatworm (fluke) [208]
Giardia (2%)	yes	protist parasite [208]
Wallemia (2%)	no	fungi commonly found in food; can cause skin infections; might transiently occur in human gut if colonized food is consumed [209]
Perkinsus (2%)	no	molluscan parasite; might transiently occur in human gut if colonized molluscan is consumed [210]
Fusarium (2%)	yes	mycotoxin-producing fungi commonly [202] found on food
Necator (2%)	yes	parasitic nematode [208]
Trepomonas (2%)	no	free-living, environmental protist whose ancestor might have been parasitic [211]
Brettanomyces (2%)	yes	yeast commonly found on fruit [204]
Trichinella (2%)	yes	parasitic nematode that can affect microbial composition of gut [212]

4.3.10 Computational resource usage of the tools

The runtime and maximum random-access memory (RAM) usage (Figure 4.6) of the tools were recorded for the non-microeukaryote and *microeukaryote_in_DB* simulated metagenomic datasets as well as one of the Cameroonian gut metagenomes—a total of ~91 million reads. All tools were run using 12 CPUs on a 256 gigabytes RAM node (Intel® Xeon®

central processing unit, E5-2650, version 3, at 2.3 gigahertz). In terms of runtimes, mTAGs had the shortest (85 minutes) followed by taxaTarget and EukDetect (102 minutes). EukDetect (1.7 GB) and taxaTarget (2 GB) had the smallest maximum memory requirement followed by mTAGs (3.2 GB). In contrast, although Metaxa2 did not have a substantially longer runtime, 130 minutes, its memory usage ranged widely from a minimum of 46 MB to a maximum of 12 GB depending on the dataset being analyzed.

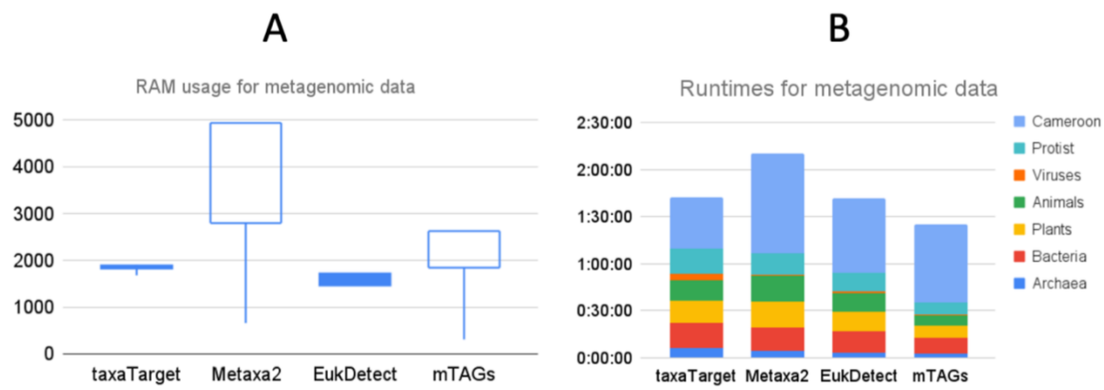

Figure 4.6 Computational efficiency of the tools when run on several metagenomic datasets. A) Maximum RAM (MB) usage. B) Runtimes (hours:minutes:seconds).

4.4 Discussion

Metagenomics is a powerful tool for exploring the complex microbial communities of bacteria, archaea, viruses, and eukaryotes that compose microbiomes. However, despite the many tools available, we are still not able to comprehensively assess the taxonomic composition of metagenomes. Here we report an important step forward for the identification of microeukaryotic species in metagenomic data by introducing taxaTarget, a fast and computationally efficient supervised learning method that learned region-specific classification thresholds from the UniRef100 database. Our results show this approach can match and often outperform other marker-gene-based tools in terms of precision and sensitivity. These gains applied for species with sequences in the database and for species which only have close relatives in the database—extending the predictive power of the scarce sequence data available for microeukaryotes. Although the use of a protein database for taxaTarget could be expected to have high sensitivity for divergent sequences, due to the redundancy of the genetic code, it was not expected to be more precise than tools using nucleotide databases (indicating that higher precision might be attained with a nucleotide database).

4.4.1 Scarce reference sequences for microeukaryotes present many challenges

Perhaps the greatest challenge for developing taxaTarget was the scarcity of reference sequence data for microeukaryotes versus their estimated species diversity [213]. It is imperative that more genomes/transcriptomes/proteomes are sequenced for eukaryotic species to increase the breadth of detectable taxa in metagenomes. As seen in the analysis of the Cameroonian gut metagenomes, although taxaTarget identified more microeukaryotes than the previous study of the dataset, including clinically relevant parasites, the biased representation of species in the databases of the tools acutely affected the results. EukDetect and taxaTarget—which use a similar marker gene database—detected an almost non-overlapping set of genera compared to Metaxa2 and mTAGs, which use databases of ribosomal subunits and COI genes. This result is somewhat expected given that the databases of Metaxa2 and mTAGs contain nearly an order of magnitude more taxa than the databases of EukDetect and taxaTarget. However, what is unexpected is that Metaxa2 and mTAGs identified a total of 44 genera and yet only shared 4 despite using very similar databases. This calls into question the validity of many of these detections and the reproducibility of results when using these markers for taxonomic classification of metagenomic data.

Scarce data is especially problematic for supervised learning because the accuracy of classification is often dependent upon the number and diversity of labelled training data. For taxaTarget, the classification thresholds of many regions were learned using five or fewer data points, preventing the use of more sophisticated machine learning approaches. Some regions lacked some or all the thresholds because there was no training data. Together, this prevented the use of neighboring windows to impute missing thresholds. However, as more data becomes available, future work should test imputation methods (a common solution for handling missing data in supervised learning [214]) as they might increase sensitivity/precision by including more regions for classification.

4.4.2 Solutions for handling scarce reference sequences

A highly effective solution, implemented in taxaTarget, for handling classification errors due to scarce data was to pad thresholds that lacked training data at broader taxonomic ranks. However, even at high threshold padding, false positives from non-microeukaryotes persisted, highlighting the need for more comprehensive training data. A persistent source of classification errors were small regions of marker genes with high sequence similarity in non-microeukaryotes. Provided sufficient depth of metagenomic sequencing, such regions might be simply filtered using outlier detection [215]; however, we often observed that each microeukaryotic taxa identified in a metagenome had less than 1,000 mapped reads—far less than 1X depth of coverage. The simple and effective solution developed for taxaTarget used a variant of the "urn problem" to detect and remove outlier marker genes. The marker genes of a taxa were treated like marbles with different colors in an urn and the number of reads mapped to the marker genes were treated as draws, with replacement, from the urn. Averaged over 100 simulations, the count of the marker gene drawn most often provided a useful baseline for the highest expected variant.

Another source of error caused by scarce data, (observed here and previously [184]) are species that have not had their full set of marker genes sequenced. For example, if *Entamoeba histolytica* has a full complement of marker genes in the database, but *Entamoeba dispar* does not, metagenomic reads, corresponding to the missing genes of *E. dispar*, could map to the orthologs in *E. histolytica* and thus get misclassified. EukDetect partially addressed this problem by filtering out species where more reads mapped to unshared versus shared marker genes with another species in the same genus. taxaTarget implemented a modified version of this solution and found it effective, although some errors persisted. EukDetect applied an additional filter (taxaTarget does not use), to further address this source of error, which removes species if the global percent identity of aligned reads across genes shared with the primary species is less than that for the primary species. It was acknowledged in [184] that this might result in only detecting the primary species of a genus, but this bias was observed for EukDetect even when using a simulated metagenome where each species had 1X depth of coverage. Future work should explore the adaptation of statistical methods for assessing ambiguity in taxonomic classification, learned from amplicon sequencing [191].

4.4.3 Insights from training and improvements

A major benefit of using a supervised learning approach was that training provided granular information about the discriminatory power of each region of the database sequences (Metaxa2, mTAGs, and EukDetect do not provide), which sometimes varied widely for adjacent regions (this information is provided for download with taxaTarget and might be useful for the study of sequence evolution). Where EukDetect removed 41 of the 255 eukaryotic BUSCO marker gene families from its database for regions conserved in bacteria, taxaTarget was able to retain all 255, by only excluding the regions that lacked discriminatory power.

Previously, the most expensive step for creating a tool like taxaTarget was the training process. To classify metagenomic reads of variable length, MetaPhyler and ROCker were trained with a minimum of three read sets of different length [192, 193]. However, we found that the longest read set was sufficient for training because the mean bit scores of the alignments were similar to those of the shorter read lengths. This resulted in a threefold reduction in training time. A potential future research area is exploring the replacement of the read mapping step of training with directly using the pairwise alignments of the sequences in the database.

4.4.4 Future work to extend taxaTarget

Future work should test the methods of taxaTarget when extended to the marker genes of other phylogroups such as bacteria, archaea, and viruses. Furthermore, the methods should be tested when extended to entire sequence databases—which might improve sensitivity and specificity. Ideally, given sufficient data, training could identify which regions of a genome or proteome are useful for taxonomic classification. However, the use of sequences other than marker genes might not be feasible without the development of additional algorithms for handling scarce data and the resulting classification errors. Lastly, despite highlighting the performance gains of data-driven methods, taxaTarget still implemented multiple user-defined thresholds, which also might be optimized or replaced with data-driven methods.

4.5 Conclusions

Our results show that the taxonomic classification of microeukaryotes in metagenomic data can be improved by utilizing data-driven methods that learn classification thresholds from the structure of an input database. These gains applied for species with sequences in the database and for species which only have close relatives in the database. The scarcity of available reference sequences still presents many challenges for the metagenomic taxonomic classification of microeukaryotes. More genomes/transcriptomes/proteomes need to be sequenced and methods for handling scarce data and the resulting classification errors need to be further explored. The implementation of these methods will help facilitate a more taxonomically comprehensive analysis of metagenomic data and expand our knowledge about roles and diversity of eukaryotes in microbial communities.

5. Assessing the use of plasmids for relating the 2020 *Salmonella enterica* serovar Newport onion outbreak to the farms implicated by the outbreak investigation

This chapter is currently being prepared for submission to BMC Genomics. It was a joint work with Hugh Rand, Kiran Javkar, Erin K. Molloy, James B. Pettengill, Arthur Pightling, Maria Hoffmann, Mihai Pop, Victor Jayeola, Steven Foley, and Yan Luo. SC, HR, YL, and KJ designed the study. SC, YL, KJ, and AP analyzed the data. SC, HR, KJ, EKM, JBP, AP, MP, SF, and YL interpreted the data. SC, HR, and YL wrote the manuscript. MH and VJ performed all sequencing and wet lab work. All authors read, provided feedback, and approved the final manuscript.

5.1 Background

In 2020, the United States Food and Drug Administration (FDA) and the Centers for Disease Control and Prevention (CDC) responded to the largest *Salmonella* outbreak in over a decade, which was caused by a strain from lineage III of *Salmonella enterica* serovar Newport (further referred to as *Salmonella* Newport). *Salmonella* Newport is one of the top five serovars contributing to the approximately 80.3 million foodborne illnesses caused by *Salmonella enterica*, globally, each year [216]. *Salmonella* Newport is polyphyletic, composed of three lineages, and is frequently associated with cattle [217] but it can colonize a wide range of animal (wild and domesticated) and plant species, providing it multiple reservoirs and multiple transmission routes to humans [218]. Furthermore, *Salmonella* Newport can persist in the environment (e.g., in manure) for months [219].

The first cases from the outbreak were reported in June of 2020 and the CDC declared the end of the outbreak in October 2020, after causing nearly 2,000 illnesses in the United States and Canada. The investigation traced the outbreak from the food history of the sick patients to red onions grown on two farms in Holtville and Bakersfield California [220]. This was the first *Salmonella* outbreak associated with bulb onions but was soon followed in 2021 by another bulb onion outbreak of *Salmonella enterica* serovar Oranienburg, lending urgency to understanding the chain of events that led to the outbreaks.

The leading hypothesis of the FDA from the on-site investigation of the onion farms was that contaminated irrigation water was used to grow the onions at the Holtville California farm [222]. Plausible sources of contamination were identified, such as sheep grazing on adjacent land and signs of animal intrusion (e.g., scat and large flocks of birds). After harvesting, the outbreak strain could have been transmitted to Bakersfield because many of the Holtville onions had been shipped to Bakersfield for packaging and distribution. Although the investigation occurred after the onions were packaged and distributed, visual observations of the packing house confirmed numerous opportunities for contamination, including signs of animal and pest intrusion, as well as food contact surfaces which had not been properly inspected, maintained, or cleaned [222].

In molecular epidemiology, if the whole genome sequencing data (WGS) of clinical and environmental isolates collected during an outbreak share a recent common ancestor—typically determined using single nucleotide polymorphisms (SNP) or core gene allele differences (cgMLST)—they are likely to be linked in the transmission chain [223]. Analysis of the WGS data with the CFSAN SNP pipeline [113] determined that the clinical isolates formed a single, highly-related clade (further referred to as the clinical clade), consistent with a single point source of contamination. However, SNP analysis also determined that none of the environmental

Salmonella enterica, from 30 serovars, collected near and on the farms (further referred to as the farm isolates), were closely related to the outbreak strain—preventing the conclusive identification of the outbreak source with the WGS data.

Here, we explored an alternative method for analyzing the whole genome sequencing data driven by the hypothesis that isolates that had recently coexisted in the same microbiome (i.e., the outbreak strain and environmental isolates) might share plasmids related by recent horizontal transfer. Horizontal transfer is the movement of genetic material between the genomes of organisms turning the tree of life into an evolutionary network [224, 225]. A common bioinformatic approach for detecting horizontal transfer events is to identify incongruences between the phylogeny of the lineages being analyzed and the phylogenies of the sequences suspected to be horizontally transferred [226]. Horizontal transfer is a fundamental source of genetic adaptation (such as antimicrobial resistance genes and metabolic pathways) to new environments and conditions for bacteria [224, 225]. Plasmids are small (generally between 1 and 200 kbp) intracellular genetic elements that can semi-independently replicate and are commonly recognized as the most impactful source of rapid horizontal transfer in microbial communities [227, 228]. Multiple plasmids, representing various incompatibility types (Inc), have been observed in *Salmonella* Newport (i.e., IncA/C, IncR, IncI1, IncN, IncH, IncF, ColE1, IncP); however, the plasmid profile is likely under-characterized given that most were identified while looking for genes conferring antimicrobial resistance [229-235].

As part of our analysis, we analyzed the pangenome—the totality of gene families found in a clade [236]—of the clinical isolates. The pangenome can be subdivided into the core genes, which are common to all the genomes and the accessory genes which are only found in a subset. Further, pangenomes can be described as closed if the number of observed gene families approaches a limit as more genomes are sampled. In contrast, in an open pangenome, the number of genes continues to grow as more genomes are sampled. A recent study of the three polyphyletic lineages of *Salmonella* Newport found that each lineage had a closed pangenome with a large number of core genes (ranging from 3,489 to 3,820), indicating the lineages underwent limited horizontal transfer [237].

We hypothesized that if the reservoir population of the clinical clade existed in the Holtville and Bakersfield farm regions, there was the possibility that it interacted with the local microbial communities via horizontally transferred plasmids. Through the analysis of the 2020 *Salmonella* Newport onion outbreak, we first explored if a highly related lineage of *Salmonella* Newport had a substantial accessory genome and plasmid diversity. Secondly, we sought to detect horizontally transferred plasmids that provided evidence the clinical and farm isolates had coexisted in the same microbiome.

5.2 Results

5.2.1 Whole genome SNP analysis of the clinical isolates

Whole genome, reference-based SNP analysis confirmed the previous findings of the FDA and CDC outbreak investigation—that the clinical isolates formed a single, low-diversity clade (further referred to as the clinical clade). The pairwise number of SNP differences ranged from 0 to 16 with a median of 0 (Figure 5.1A). For reference, the median assembly length for the clinical isolates was 4.8 Mbp. The phylogeny constructed from the SNP results revealed a single large polytomy which precluded any observations about the clustering of the clinical isolates by collection date or geographic location.

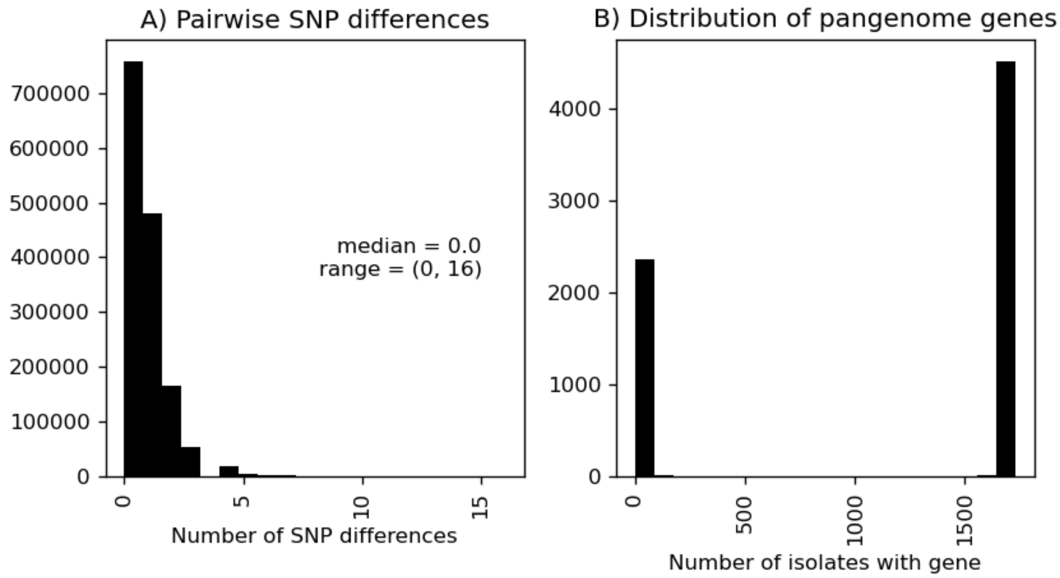

Figure 5.1 Genetic characteristics of the clinical isolates
A) The median number of pairwise SNPs. B) A histogram showing the number of isolates each of the 6,976 pangenome genes occurs in.

To assess if the clinical clade was associated with a geographical location, the previous SNP analysis was extended to include all closely related (within 1,000 cgMLST alleles of the clinical isolates) environmental isolates from the NCBI Pathogen Detection database (Figure 5.2). The most closely related isolates were mainly from California, New Mexico, and Mexico. The clinical isolates formed a subclade within the SNP phylogeny with 10 environmental isolates from California (seven isolated from almonds, one from parsley, two from environmental swabs) and one from Washington state (isolated from pistachios likely grown in California [238]). These isolates had been collected between 2010 and 2017 and were between 8 and 39 SNPs distant from the clinical isolates. Sister to this subclade were environmental isolates from California (comminuted beef), New Mexico (environmental swabs from an unknown source), and Mexico (mostly water samples from rivers, canals, and a reservoir, with one from a chicken caecum) collected between 2014 and 2020.

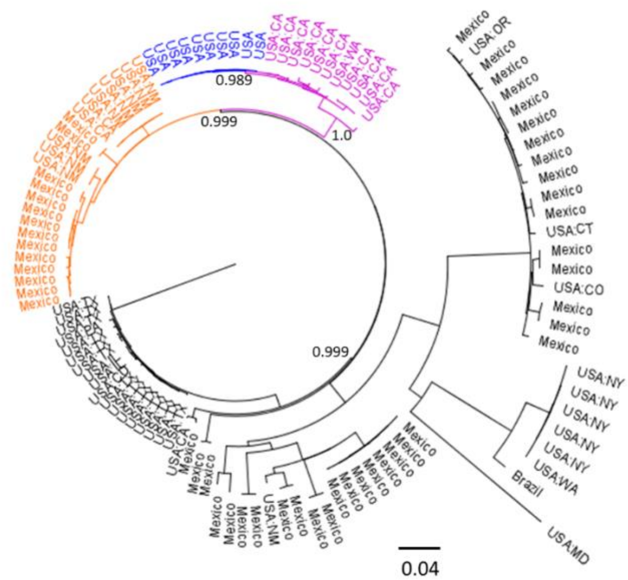

Figure 5.2 Geography of the phylogenetic neighborhood of the clinical clade
The SNP-based maximum likelihood phylogeny contains all environmental isolates identified in the NCBI Pathogen Detection database within 1000 cgMLST alleles of the clinical clade. Only select bootstrap values that highlight the separation between the clinical and environmental isolates are shown for clarity. The clinical clade (blue) is nested within a larger clade (magenta) of ten isolates collected from California and one isolate from Washington, collected between 2010 and 2017. The sister clade (orange) contains isolates from California, New Mexico, and Mexico collected between 2014 and 2020. The phylogeny was inferred using GARLI (with 1,000 bootstraps) with the SNP matrix generated by the CFSAN SNP Pipeline.

5.2.2 The pangenome of the clinical clade had a substantial accessory genome

The core genome of the clinical clade was highly conserved, consistent with the low diversity observed in the SNP analysis. There were 4,399 core genes (genes occurring in 99% of isolates) which was near the median number of predicted genes per genome (4,512 genes). However, despite a large and conserved core genome, the accessory genome still contained a substantial number of genes (2,577). These accessory genes were sparsely distributed, mostly occurring in 6% or less of the isolates (Figure 5.1B).

5.2.3 Plasmids of the clinical and farm isolates

Pangenome analysis of the clinical clade revealed a substantial but sparsely distributed accessory genome, indicative of horizontal transfer. We next sought to assess how many plasmids were contained in the clinical and farm isolates. For the clinical isolates, 1,814 putative extrachromosomal plasmid contigs were identified, belonging to 20 known plasmid types (Table 5.1), as well as 373 putative plasmid contigs of unknown type. The plasmids of known type accounted for 1,590 genes of the clinical clade pangenome—16 core and 1,574 accessory genes (≥64% of the accessory genes)—and ranged in size from approximately 1.6 kbp to 85 kbp long. The most observed plasmid type was the IncFII(S), occurring in all clinical isolates. Other plasmid types occurred infrequently. The second most abundant plasmid type, IncI1-I(Gamma), was only identified in 27 (1.6%) clinical isolates. No clustering of the plasmids was observed in the whole genome SNP phylogeny due to it being a large, unresolved polytomy. Additionally, no plasmid was found to be significantly associated with hospitalization.

Table 5.1 Summary of the plasmids found in the clinical and farm isolates.

Plasmid type	Number of clinical isolates with plasmid	Number of Holtville isolates with plasmid	Number of Bakersfield isolates with plasmid	Median contig length (bp)	Median number of genes
IncFII(S)	1728	3	7	72766	76
Unclassified	373	235	42	6170	7
Col440I	3	60	8	4322	5
IncI1-I(Gamma)	27	13	3	85161	93
IncQ1	0	20	6	8259	10
IncFIB(pB171)	0	25	0	28614	33
Col(pHAD28)	1	12	4	4751	7
IncFII	1	15	0	70562	85
ColpVC	12	4	0	2223	1
IncX3(pEC14)	0	14	0	38955	51
IncFII(pCRY)	0	3	10	49311	62

Col(BS512)	12	0	0	2216	2
ColRNAI	0	1	10	9712	11
IncI2(Delta)	6	0	0	58474	75
IncI(Gamma)	4	0	0	84482	95
Col(MG828)	4	0	0	1672	1
IncFII(Yp)	0	4	0	171753	190
IncR	0	4	0	13431	15
pXuzhou21	3	0	0	40119	54
IncL	3	0	0	59108	81
Col156	2	0	0	4937	5
IncX4	2	0	0	30161	44
Col8282	1	0	0	4218	3
IncB/O/K/Z	1	0	0	21481	25
IncN	1	0	0	38347	49
IncX1	1	0	0	40944	51
pSL483	1	0	0	38560	53
IncM1	1	0	0	58540	76
IncFII(SARC14)	0	1	0	31141	24
Total	2187	414	90		

The 512 farm isolates, from 30 different *Salmonella* serotypes, contained 227 putative extrachromosomal plasmid contigs belonging to 14 known plasmid types and 277 putative plasmid contigs of unknown type (Table 5.1). No plasmid type was common to all isolates and 346 isolates had no identified plasmid contigs. The most observed plasmid types were the Col440I (68 isolates) and the IncQ1 (26 isolates).

At a broad level, six plasmid types were observed in both the clinical and farm isolates: IncFII(S), IncFII, IncI1-I(Gamma), ColpVC, Col440I, and Col(pHAD28). There were also many uncharacterized plasmids observed in both.

5.2.4 Analysis of plasmids sharing high similarity in the the clinical and farm isolates

For our analysis, we sought to identify plasmids that might have undergone recent horizontal transfer between the clinical and farm isolates. To do this, we restricted our analysis to clinical and farm plasmids sharing high similarity (≥95% identity and 90% alignment coverage) i.e., those more likely to be related by recent horizontal transfer. Amongst the six plasmid types observed in both the clinical and farm isolates, as well as the uncharacterized plasmids, there were 13 clinical isolates containing 14 plasmids with high similarity to 17 farm plasmids in 8 isolates (Figure 5.3). The highly similar clinical and farm plasmids belonged to the Col440I, ColpVC, and IncI1-I(Gamma)) plasmid types as well as one uncharacterized plasmid type (Table 5.2). The eight farm isolates with these plasmids had been collected from 2 sampling sites separated by 409 km (254 miles): 1) water samples from the New River in Seeley, California (about 20 miles West of Holtville); 2) soil samples near an irrigation filling station next to the Bakersfield onion farm. Five isolates (four *Salmonella* Corvallis and one *Salmonella* Liverpool) were collected from the first site and each of them carried two or three high similarity plasmids. The other three isolates, two from *Salmonella* Idikan and one from *Salmonella* Typhimurium were collected from the second site and only carried one high similarity plasmid per isolate. Amongst the clinical isolates, all carried a single plasmid with high similarity to a farm plasmid except isolate SRR12424118 which carried two.

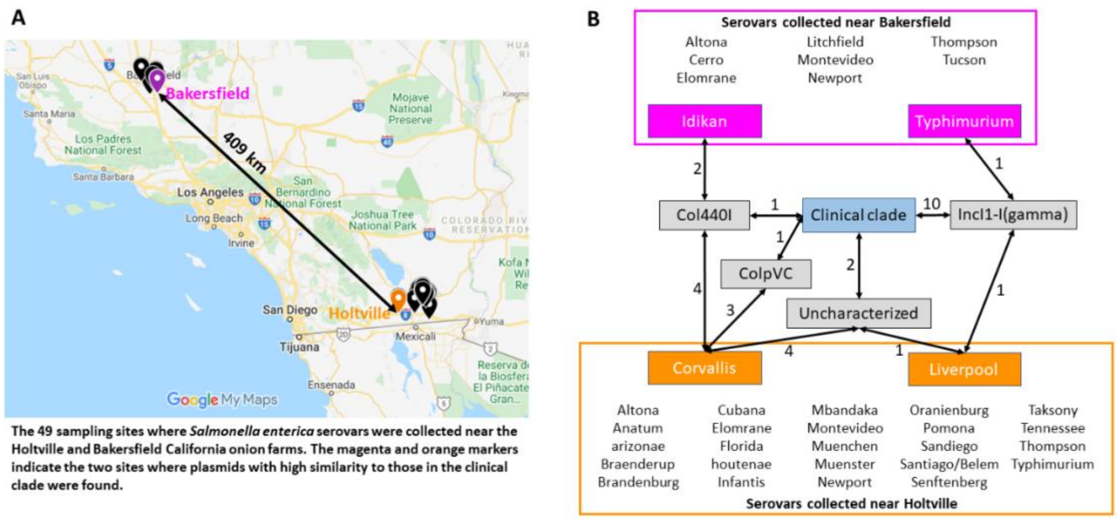

Figure 5.3 High similarity plasmids shared by the clinical and farm isolates

A) The map (made in Google My Maps) shows the 49 sampling sites where farm isolates were recovered and the two sites, near Holtville (orange) and Bakersfield (magenta), where high similarity plasmids to those in the clinical clade were found. B) The graph shows the relationship of similar plasmids in the clinical clade and farm isolates. The magenta and orange boxes list the *Salmonella enterica* serovars that were collected from the Bakersfield and Holtville regions, respectively. Edges connect the serovar or clinical clade to the plasmid they share. Edge labels indicate the number of isolates with the shared plasmid. The grey ellipses indicate the plasmid type that is shared.

Table 5.2 Summary of metadata for the highly similar clinical and farm plasmids.

Plasmid type	Number of isolates with plasmid (clinical, Holtville, Bakersfield)			Number of highly similar plasmids (clinical, Holtville, Bakersfield)			Geographic extent of NCBI best hits	Environmental isolation sources of NCBI best hits	Temporal range of NCBI best hits	Bacterial host range (genera) of NCBI best hits
IncFII(S)*	1728	3	7	1728	3*	7*	United States, Mexico	River water, soil, almonds, pistachios, chicken	2010 to 2020	*Salmonella*
IncI1-I(Gamma)	27	13	3	10	1	1	United States, South Korea, Canada, United Kingdom, Denmark, Switzerland,	Cow, sheep, dog, pig, chicken, catfish, horse, iguana, lettuce, soil	2002 to 2022	*Escherichia*, *Salmonella*, Shigella
ColpVC	12	3	0	1	4	0	United States, Canada	Chicken, turkey, pig	2007 to 2020	*Salmonella*
Col440I	3	60	8	1	4	2	United States, Germany, Mexico, Venezuela, Ecuador	Chicken, cow, pig, papaya, river water	2013 to 2020	*Escherichia*, *Salmonella*
Unclassified	373	235	42	2	5	0	United States, Australia, United Kingdom, India, Japan	Dog, cow, pig, chicken, wastewater	2014 to 2021	*Escherichia*, *Salmonella*, Klebsiella

Next, we sought to include closely related plasmids from the NCBI for phylogenetic analysis. The 14 clinical plasmids (sharing high similarity with the 17 farm plasmids) were used to recruit the best BLAST hits from the NCBI Pathogen Detection and Nucleotide databases. The best BLAST hits were filtered for plasmids from environmental isolates that shared as high of sequence similarity as that between the clinical and corresponding farm plasmids. We restricted our analysis to environmental isolates to increase the chance of discovering signal from geographic and isolation sources. With the clinical, farm, and NCBI plasmid sequences, multiple sequence alignments were built, and phylogenetic analysis was performed. In some cases, the phylogenies lacked resolution, i.e., formed large polytomies, and so graphs (referred to here as mutation graphs) were built to visualize the mutational similarities and differences between the plasmid sequences. The mutation graphs presented are not intended as phylogenetic analyses [239]

For the phylogenetic analysis of the plasmids, we hypothesized that clinical and farm plasmids related by recent horizontal transfer would belong to the same subclade within the phylogeny. However, it was uncertain how close they would cluster in the phylogeny because plasmids can be rapidly transmitted between species and the transmission chain between clinical and farm isolates might have involved intermediate species [240-242]. A brief description of each plasmid as well as the results from the multiple sequence alignments, phylogenies, and mutation graphs are described in the paragraphs below.

5.2.4.1 IncI1-I(gamma) plasmid

IncI1 plasmids have been widely observed in *Enterobacteriaceae* and isolated from many animals. This plasmid type often confers antimicrobial resistance and/or colicin production

[243]. Those analyzed here approximately had an average length 85 kbp and 90 predicted genes. Most of the genes could not be functionally annotated (72 genes), however, genes with predicted functions were involved with the Type II secretion system, plasmid segregation and partitioning, bacterial outer membrane adhesion, pilus formation, DNA polymerase IV, and a Colicin-Ia operon.

The IncI1-I(gamma) plasmid type occurred in 27 clinical isolates and 16 farm isolates, but only ten clinical and two farm plasmids (1 Holtville *Salmonella* Liverpool isolate and 1 Bakersfield *Salmonella* Typhimurium isolate) shared high similarity. The whole plasmids could not be aligned due to different gene content and order, so the concatenated core genes were used for the multiple sequence alignment. Analysis of the multiple sequence alignment revealed that the Bakersfield and Holtville plasmids were more distant from each other (276 SNPs) than to any clinical plasmid. Further, the clinical plasmids were more similar to the Holtville (median = 162 SNPs) than the Bakersfield (median = 230 SNPs) plasmids. Additionally, the median SNP distance between the clinical isolates (146 SNPs) was similar to the distance between the clinical and Holtville farm plasmids (155 SNPs). According to metadata held internally by the FDA, the clinical isolates carrying these plasmids had been collected from patients over the duration of a month in multiple states in the USA as well as Canada.

The phylogeny of the clinical and farm plasmids and the best BLAST hits from the NCBI showed that eight of the clinical plasmids belonged to a subclade with the two farm isolates (Figure 5.4). This subclade also contained plasmids from four *Escherichia coli*, one *Escherichia fergusonii*, one *Shigella flexneri*, one *Salmonella* Newport, and one *Salmonella* Derby isolates. These isolates were collected from 2002 to 2019, from the United States, Canada, the United Kingdom, South Korea, and Denmark, from cows, sheep, dogs, pigs, and soil samples.

Figure 5.4 Phylogeny of the IncI1-I(Gamma) plasmids
A) The unrooted maximum likelihood phylogeny of the IncI1-I(Gamma) plasmid (branch lengths are not shown). The phylogeny was built from the multiple sequence alignment of the concatenated core genes (33 genes and 24,173 bp). There were 23,180 invariant sites in the multiple sequence alignment. The bootstrap values are provided as internal node labels. Of the 27 clinical isolates and 16 farm isolates that had this plasmid, only ten clinical isolates (blue) and two farm isolates (orange) shared high similarity.

5.2.4.2 Uncharacterized plasmid

There were 373 and 277 putative plasmid contigs that could not be classified as a known plasmid type in the clinical and farm isolates, respectively. Amongst these, two clinical and five farm plasmid sequences shared high sequence similarity. These sequences corresponded to a cryptic plasmid previously observed in species across the *Enterobacteriaceae* family [244-246]. The sequences analyzed here were approximately 4,200 bp long with six predicted genes including a *repB* replication gene, a *mobQ* relaxase, a conjugal transfer gene, and three hypothetical genes.

The two clinical plasmids differed by 6 SNPs and had been collected from two patients in Oregon and Missouri separated by 13 days. All five of the farm plasmids were collected on the same day at the Holtville collection site, with four in *Salmonella* Corvallis isolates and one in a *Salmonella* Liverpool isolate. The four Holtville *Salmonella* Corvallis plasmids differed from each other by 2 SNPs and the *Salmonella* Liverpool plasmid by 62 SNPs. Sixteen highly similar plasmids were found in the NCBI Nucleotide and Pathogen databases for which there was no consistent name or type.

The phylogeny showed that the two clinical plasmids and four Holtville *Salmonella* Corvallis plasmids belonged to sister subclades within a larger subclade. These clinical and farm plasmids differed by 29 to 41 SNPs (Figure 5.5). The subclade of the clinical plasmids also contained plasmids from three *E. coli* and one *Salmonella* Heidelberg isolates collected from several animals (dog, cow, pig, chicken) in the United States between 2016 to 2021. The subclade of the four Holtville plasmids contained one *Salmonella* Typhimurium isolate collected from a pig in Australia in 2014.

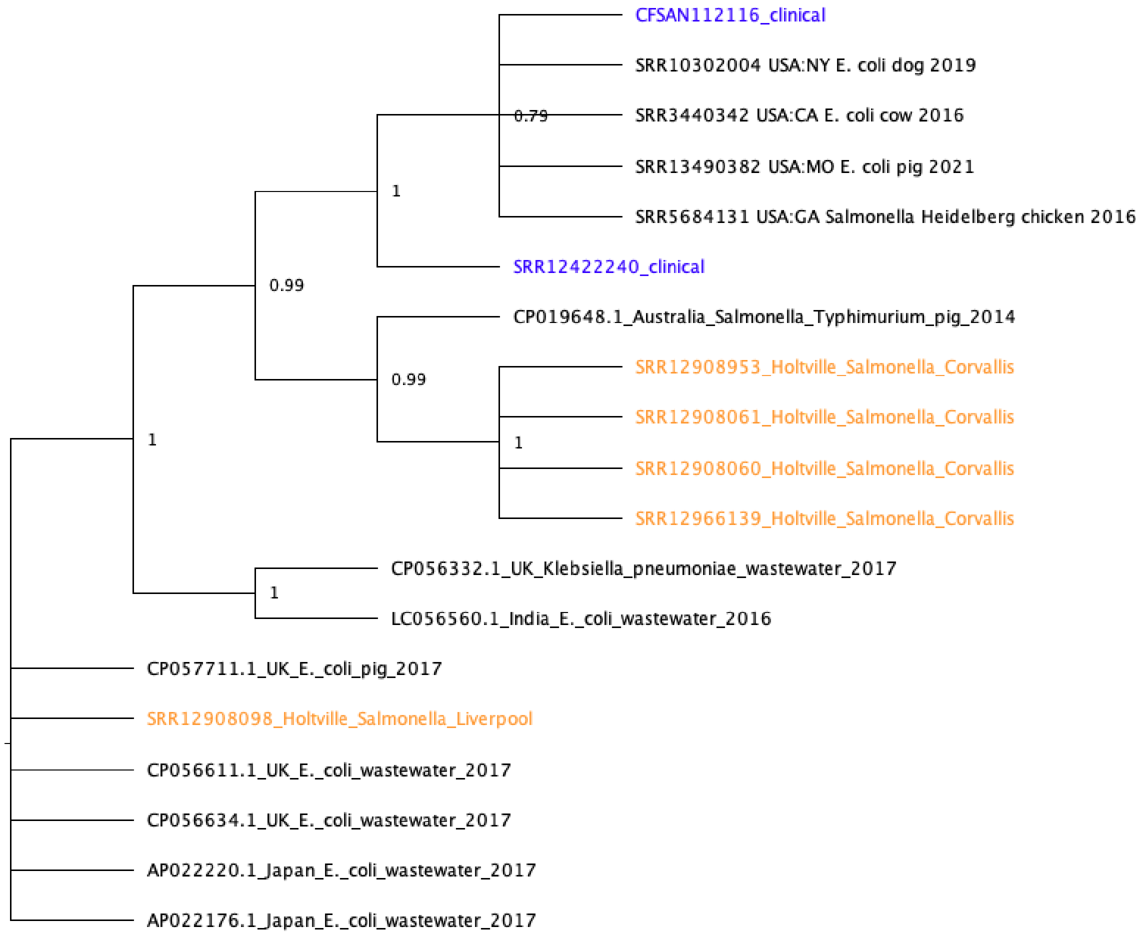

Figure 5.5 Phylogeny of the uncharacterized plasmids
The unrooted maximum likelihood phylogeny of the uncharacterized plasmid (branch lengths are not shown). The phylogeny was built using the multiple sequence alignment of the whole plasmids (4,201 bp) which had 3,956 invariant sites. The Holtville isolates are in orange and the clinical isolates are in blue. Four of five Holtville isolates formed a clade, each differing by 2 SNPs from the rest. These Holtville isolates were 2 SNPs different from a *Salmonella* Typhimurium isolate collected from a pig in Australia in 2014. The clinical plasmids (which differed from each other by 6 SNPs) were in a sister clade to the four Holtville plasmids, differing by 29 to 41 SNPs. Bootstrap values are provided as internal node labels.

5.2.4.3 Col440I plasmid

The Col440I plasmid has been observed in species across the *Enterobacteriaceae* family [244, 245] and those analyzed here had three predicted genes: *qnrB* (quinolone resistance gene [247]), a phage shock protein transcription activator (*psp* gene), and a pentapeptide repeat protein. The plasmid type was only found in three clinical isolates but was the most frequently observed plasmid type in the farm isolates, occurring in 68. However, only one clinical and six farm plasmids shared high similarity. For the highly similar farm plasmids, four were found in *Salmonella* Corvallis isolates at the collection site near the Holtville farm and two were found in *Salmonella* Idikan isolates at the collection site near the Bakersfield farm.

In the NCBI Nucleotide and Pathogen databases there were 19 highly similar plasmids. These had been collected from 2013 to 2020 from diverse isolation sources (chickens, cows, pigs, papaya, river water, wastewater) and geographic locations (USA, Germany, Canada, Mexico, Venezuela, Ecuador), as well as multiple *Salmonella enterica* serovars and E. coli.

The phylogeny formed a single, unresolved polytomy so a mutation graph was made to visualize the differences between the sequences (Figure 5.6). The mutation graph showed that there were only 3 SNP differences between the clinical plasmid and the other plasmids, which were all identical.

Figure 5.6 The Col440I plasmid mutation graph

Col440I plasmid mutation graph using the clinical plasmid as the reference. For example, A1922 means a plasmid was identical to the clinical plasmid except a mutation to A at loci 1922. The graph was built using the multiple sequence alignment of the plasmids (2,699 bp) which had 2,696 invariant sites. This graph shows mutational differences and does not necessarily represent phylogenetic relationships.

5.2.4.4 ColpVC plasmid

The ColpVC plasmid is a cryptic plasmid [248] that has been observed in species across the *Enterobacteriaceae* family [244, 245]. Those analyzed here had one replication gene (*rep*, pfam01446) and one hypothetical gene. The plasmid type was observed in 12 clinical isolates and 4 *Salmonella* Corvallis farm isolates from the Holtville collection site. Only one of the clinical ColpVC plasmids was highly similar to the 4 farm plasmids, which differed by 1 or 3 SNPs. The four farm plasmids had been collected on the same day, from the same site and serovar, but differed by 0 to 4 SNPs.

Within the NCBI Nucleotide and Pathogen databases there were 100 highly similar plasmids found. These had been collected from the USA, UK, and Canada; from 1999 to 2021; mostly from chickens and turkeys; mainly from *Salmonella enterica*—mostly serovars Reading, Kentucky, and Enteritidis—as well as *E. coli*.

The phylogeny formed a single, unresolved polytomy so a mutation graph was made to visualize the differences between the sequences (Figure 5.7). The mutation graph revealed 57 plasmids that were identical to the clinical plasmid. These had been collected from the USA; chickens or turkeys; various *Salmonella enterica* serovars—mostly Reading, Kentucky, and Enteritidis; and from 2007 to 2020. None of the plasmids from the NCBI were identical to the farm plasmids, but several shared one or two of the same SNP differences with the clinical plasmid.

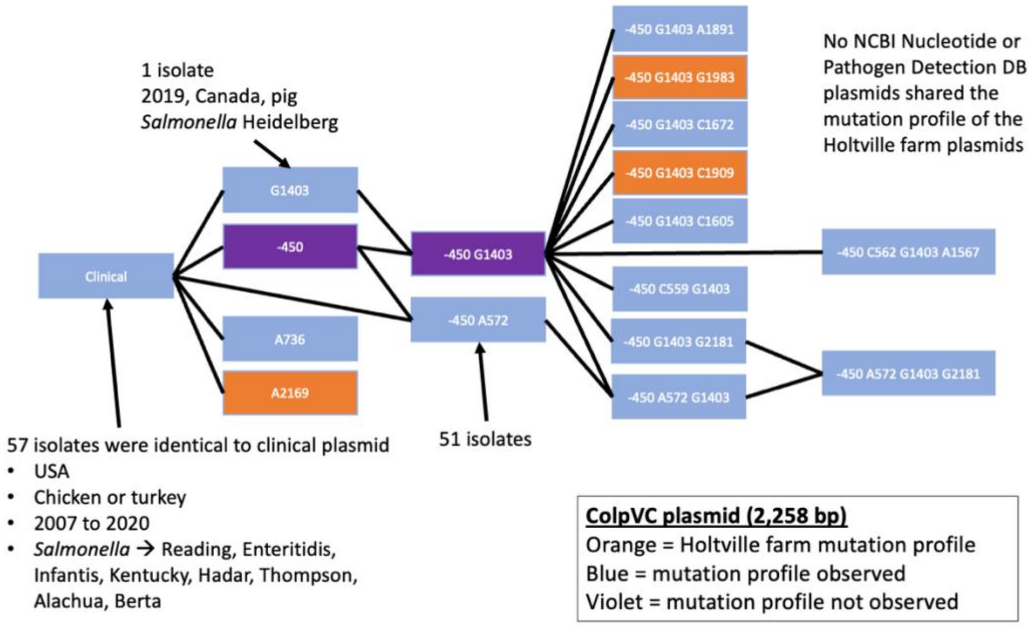

Figure 5.7 The ColpVC plasmid mutation graph

ColpVC plasmid mutation graph using the clinical plasmid as the reference. For example, G1403 means a plasmid was identical to the clinical plasmid except a mutation to G at loci 1403 (a "-" mutation indicates a deletion). The graph was built using the multiple sequence alignment of the plasmids (2,258 bp) which had 2,237 invariant sites. This graph shows mutational differences and does not necessarily represent phylogenetic relationships.

5.2.5 Analysis of the IncFII(S) plasmid

Although no high similarity IncFII(S) plasmids were found in the farm isolates, we performed a more granular version of the previous analysis for this plasmid type because it was carried by all clinical isolates. The IncFII(S) plasmid was predicted to be host restricted to the *Salmonella* genus [244, 245]. The clinical versions of this plasmid were approximately 73 kbp long with 77 genes coding for the Type-F conjugative transfer system, plasmid replication and persistence (e.g., *ccdA*/*ccdB* toxin-antitoxin system), as well as a *saf* fimbrial operon strongly correlated with increased virulence in humans [249]. Only two annotated genes appeared to confer metabolic functions to the bacterial host i.e., ammonia monooxygenase and succinate dehydrogenase flavoprotein.

Although the IncFII(S) had no full length, high similarity matches between the clinical and farm plasmids, further analysis revealed 10 farm isolates with identical genes for the pMLST profile (FIC_5, FIIS_1, FIIY_10). Three were *Salmonella* arizonae isolates collected near the Holtville farm which shared 29 genes (out of a median of 77 genes per plasmid) with the clinical plasmid. The other seven isolates were *Salmonella* Typhimurium collected from Bakersfield which shared four genes with the clinical plasmid.

Within the NCBI Pathogen Detection and Nucleotide databases, there were thirty full length, highly similar instances found in environmental isolates (25 *Salmonella* Newport and 5 *Salmonella* Javiana). The five *Salmonella* Javiana isolates had been collected from leafy greens and poultry—the only connection to an animal source. The thirty environmental isolates were all from California or Mexico, except one from Arizona and two from Washington. Twenty of the environmental isolates had been identified previously as being 8 to 39 SNPs different than the clinical isolates by the whole genome SNP phylogeny.

The phylogeny of the IncFII(S) plasmid mostly formed a single unresolved polytomy except for three isolates which branched into another polytomy. As such, a mutation graph was constructed to better observe the differences between the isolates (Figure 5.8). The most similar plasmids (having one or zero SNP difference) were from the same California isolates that had formed a subclade with the clinical isolates in the SNP phylogeny.

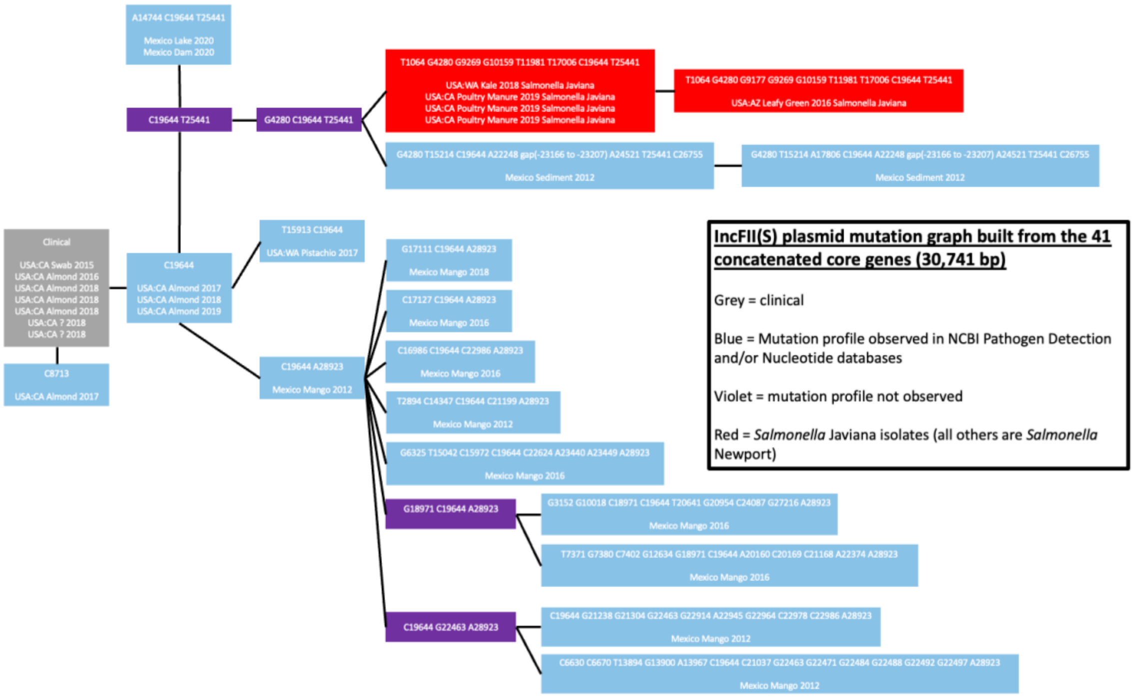

Figure 5.8 The IncFII(S) plasmid mutation graph
The IncFII(S) plasmid mutation graph using the clinical plasmid as the reference. The mutation graph was built from the multiple sequence alignment of the 41 concatenated core genes (30,741 bp) which had 29,882 invariant sites. For example, C8713 indicates a plasmid was identical to the clinical plasmid except for a mutation to C at loci 8713 (a "-" mutation indicates a deletion). This graph shows mutational differences and does not necessarily represent phylogenetic relationships.

5.2.6 Assessing the frequency of observing the highly similar plasmids

Previously 14 clinical plasmids that were highly similar to 17 farm plasmids from four plasmid types had been identified (Table 5.2). Analysis of the phylogenies, mutation graphs, and multiple sequence alignments for these plasmids supported that most of the clinical and farm plasmids shared a recent common ancestor. The only exception were two clinical IncI1-I(Gamma) plasmids which did not fall in the same subclade as the farm plasmids (Figure 5.4). We sought to assess how frequently this number and combination of highly similar plasmids could be found in random samples of environmental *Salmonella enterica*. The null hypothesis for this experiment was that just as many highly similar ColpVC, Col440I, and IncI1-I(Gamma) plasmids, as found in the outbreak isolates, could be identified in randomly sampled environmental *Salmonella enterica* isolates. For this analysis, we did not include the uncharacterized plasmids as we could not confidently identify all instances in the clinical and farm isolates. Also, the two clinical IncI1-I(Gamma) plasmids which did not fall in the same subclade as the farm plasmids were removed. This left 10 clinical and 11 farm plasmids from three plasmid types for further consideration.

To test the null hypothesis, 516 isolates (the number of farm isolates) were randomly sampled from the 95,284 environmental *Salmonella enterica* in the NCBI Pathogen Detection database (excluding those from the outbreak). All the clinical ColpVC, Col440I, and IncI1-I(Gamma) plasmids—not just those that were highly similar to the farm isolate plasmids—were then aligned to the isolate assemblies. Assemblies with plasmids sharing as high of sequence similarity as the highly similar clinical and farm plasmids were counted (Table 5.3). This was repeated 10,000 times and the frequency of observing the plasmids was recorded.

Table 5.3 Summary of highly similar plasmids used for statistical analysis.

Plasmid type of highly similar clinical and farm plasmids	Total number of clinical plasmids	Number of highly similar clinical and farm plasmids (clinical, farm)	Minimum percent identity between full length alignments of highly similar clinical and farm plasmids
ColpVC	12	5 (1, 4)	99.89%
Col440I	3	7 (1, 6)	99.87%
IncI1-I(Gamma)	27	10 (8, 2)	98.55%

First, the frequency of observing individual plasmids and plasmid types was assessed (Figure 5.9 A-D). The frequency of observing at least one highly similar plasmid of any type was

100%. The frequency of observing at least one highly similar ColpVC, Col440I, or IncI1-I(Gamma) plasmid was 99.7%, 85.1%, and 20.7%, respectively. However, the frequency of observing the outbreak counts for the ColpVC, Col440I, or IncI1-I(Gamma) plasmid types was 99.1%, 0%, and 2.8%, respectively. Additionally, the frequency of observing 3 or more isolates with multiple highly similar plasmids, as seen in the outbreak, was 0%.

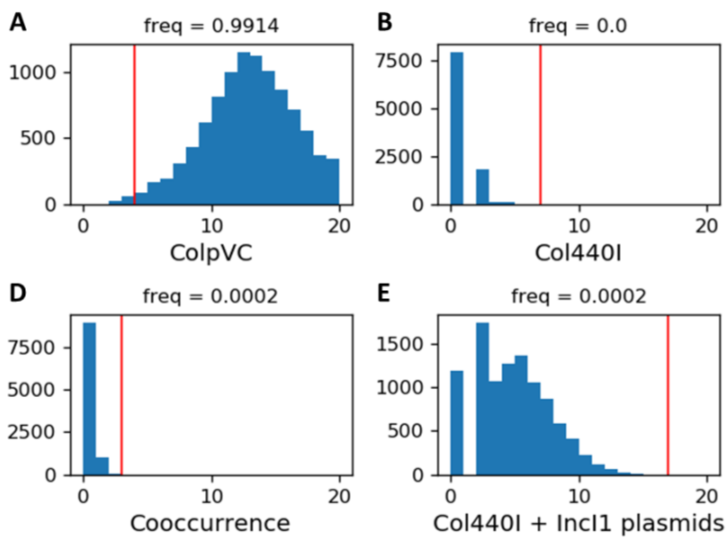

Figure 5.9 Frequency of observing highly similar plasmids in environmental *Salmonella enterica*
Frequency of observing highly similar plasmids in 10,000 random samples of 516 environmental *Salmonella enterica* isolates. The histograms show the number of A) highly similar ColpVC plasmids; B) highly similar Col440I plasmids; C) highly similar IncI1-I(Gamma) plasmids; D) isolates with multiple highly similar plasmids; E) the sum of highly similar Col440I and IncI1-I(Gamma) plasmids. The red vertical lines show the numbers observed in the outbreak. Frequencies show the proportion of 10,000 samples with at least the same number of occurrences as the outbreak.

Second, the frequency of observing plasmid profiles similar to the outbreak in the random samples was assessed (Figure 5.9E). The frequency of observing one of each plasmid type was 18%, and two of each type was 0.6%. The frequency of observing at least the same sum of highly similar Col440I and IncI1-I(Gamma) plasmids as in the outbreak was 0.01%. The

ColpVC plasmids were excluded from these sums because the previous analysis had found them to be commonly observed and thus not informative here.

5.3 Discussion

The epidemiological investigation of the 2020 *Salmonella* Newport onion outbreak strongly implicated two onion farms in Holtville and Bakersfield California as the likely source of the outbreak [220, 222]. SNP analysis provided support that the clinical isolates originated from California because the most closely related environmental isolates from the NCBI Pathogen Detection database were mainly collected from California. Further, these isolates suggested a specific link to central California because many were collected from almonds. Nearly ~100% of the almonds grown in the United states come from California and ~75% are produced in five counties in central California, i.e., Stanislaus, Fresno, Kern (where Bakersfield is located), Merced and Madera [250, 251].

We had hypothesized that if the outbreak strain had existed in the Holtville and Bakersfield farm regions, it might have interacted with the local microbial communities via horizontally transferred plasmids. Our analysis identified highly similar plasmids in the clinical and farm isolates that were likely related by recent horizontal transfer. Statistical analysis further suggested that it was unlikely to observe the number and combination of highly similar plasmids in random samples of environmental *Salmonella enterica*. Together, our results support the findings of the epidemiological investigation and suggest that the outbreak strain had come from the implicated farm regions.

A limitation of our study was that the farm isolates were the only representatives of the farm microbiomes. Although plasmids can be transferred between *Salmonella enterica* serovars directly, the local microbiota often serves as a reservoir for plasmids and can be an important intermediary because conjugation is dependent upon the proximity of donor and recipient cells as well as the density of cells [240-242]. Therefore, ideally other species of *Enterobacteriaceae* would have been collected as well as metagenomic data to increase the chance of detecting horizontally transferred plasmids. Additionally, it would have been informative to sample the gut microbiomes of local animals like cattle, which are thought to be a reservoir of *Salmonella* Newport [217], because the gut microbiome presents ideal conditions for the horizontal transfer and persistence of plasmids [240-242]. Ideally, there would have also been Enterobacteriaceae isolates and metagenomes from the gut microbiomes of patients to differentiate plasmids that were acquired there from other sources.

The analysis of the plasmid multiple sequence alignments, phylogenies, and mutation graphs highlighted important gaps in our knowledge about specific plasmid types—especially at very granular levels of microevolution. For example, identical instances of the clinical ColpVC plasmid (Figure 5.7) could be found over the last 15 years. And yet the farm ColpVC plasmids differed by 0 to 4 SNPs despite being collected from the same sampling site on the same day. Similarly, the two clinical instances of the uncharacterized plasmid (Figure 5.5) differed by 6 SNPs despite being collected within 13 days of each other (from different patients in different states). And yet, within the plasmid phylogeny the clinical isolates were separated by environmental isolates which had been collected up to six years before. Potentially, the clinical instances represent different lineages of the plasmid and were independently acquired, but the four *Salmonella* Corvallis farm isolates showed a similar pattern—each differed from the others by two SNPs—despite being collected from the same sampling site on the same day. An alternative explanation is plasmid heterozygosity in the clinical and farm isolates i.e., copies of a plasmid with sequence variations within a single cell or population [252]. It is also possible that the plasmids might have a high mutation rate but the rate at which they are fixed in the population is low [253]. Adding further complexity, the *Salmonella* Liverpool instance of the uncharacterized plasmid differed from the *Salmonella* Corvallis instances by 62 SNPs despite being collected from the same sampling site and the same day, indicating potential barriers to conjugation despite proximity.

Another challenge when analyzing the plasmids was that they showed signs of high promiscuity e.g., isolated from multiple species within *Enterobacteriaceae*, isolated from multiple animal and environmental sources, an international distribution, all within the last 20 years. This prevented the extraction of information about geographic location, isolation source, the time of transfer, and if there had been direct transfer between the clinical and farm *Salmonella* serovars or if there were intermediary species.

Together, these issues highlight that more work is needed to characterize the molecular evolution of specific plasmid types. Additionally, more environmental sampling is needed to provide a better understanding of the distribution, population dynamics, and transmission dynamics of plasmids in the environment. Essential to this effort are comprehensive sequence databases and quality metadata. Additional sequence data and metadata might have revealed more details about the plasmid phylogenies e.g., the transmission chain between bacterial species, the geographic source, the isolation source, timeframe of transmission. A general implication is that care needs to be exercised when removing "redundant" genome sequences from databases based upon core genome similarity as accessory sequences might be lost.

At a broader level, identifying 20 plasmid types in a very homogeneous population of *Salmonella* Newport lineage III was surprising given that our literature review indicated that only 11 plasmid types were previously observed in *Salmonella* Newport [229-235]. This indicates a much greater diversity of plasmids in *Salmonella* Newport than currently reported and warrants further investigation. Importantly, because many of the observed plasmid types were rare, future studies should examine large sets of isolates. The large diversity of plasmids might be explained by the dynamics of the outbreak, e.g., a population that rapidly grew and changed environments. For example, horizontal transfer can increase as a result of host transition e.g., transitioning from the environment or reservoir host to humans [254]. Future work should further explore how outbreaks affect plasmid diversity within pathogen populations. Additionally, although we focused on plasmids because they constituted the majority of the accessory genome, future studies should explore other types of mobile genetic elements such as phages and transposons. Studies should also continue to assess the burden of evidence needed to relate a pathogen to an environmental microbiome using horizontal transfer events.

More work is also needed to improve the comprehensiveness of plasmid classification. Commonly used plasmid typing methods (e.g., MOB and incompatibility typing) can fail to classify 50% of the plasmids in highly curated datasets (e.g., RefSeq plasmids) much less environmental samples [221]. Further, current typing methods are most comprehensive for Enterobacteriaceae, but we still observed hundreds of putative plasmid contigs that could not be classified. Related to classification, it can be difficult to assess the evolutionary relatedness of plasmid sequences [221, 255]. For example, the three pMLST genes of the IncFII(S) plasmid were identical in the clinical isolates as well as several farm and environmental isolates going back a decade. When using the replicon-based pMLST system [256], these plasmids would be considered highly related. However, the amount of shared gene cargo for these plasmids ranged from 5% to 100%—in all cases, the shared gene cargo had high sequence similarity. Here, using the proportion of shared genes as a distance metric [255] would find the plasmids with less shared gene cargo as more divergent. Methods are needed to reconcile the molecular evolution of core genes and cargo genes when analyzing plasmids.

5.4 Conclusion

The analysis of highly similar plasmids, found in clinical and environmental isolates, from the 2020 *Salmonella* Newport onion outbreak suggested they were related by recent horizontal transfer. However, the seemingly high promiscuity of the plasmids prevented the extraction of information about geographic location, isolation source, the time of transfer, and if there had been direct transfer between the clinical and farm *Salmonella* serovars or if there were intermediary species. Nonetheless, statistical analysis suggested that the number and combination of highly similar plasmids was unlikely to be found in random samples of environmental *Salmonella enterica*. Together out findings support the finding of the FDA and CDC investigation that the outbreak strain had likely originated from two onion farms in California. Horizontally transferred plasmids can potentially provide information about microbial interactions, connecting lineages to environments and microbiomes. Such analyses will add a new dimension to source tracking investigations but are dependent upon detailed and accurate metadata, more extensive environmental sampling, and a better understanding of plasmid molecular evolution.

5.5 Methods

5.5.1 Data and Sequencing

The data consisted of MiSeq Illumina short read data from clinical and environmental sources. The clinical clade consisted of 1,728 clinical *Salmonella enterica* (serovar Newport) isolates collected during the 2020 onion outbreak (June to October) in the United States (1,173 isolates) and Canada (555 isolates) [257]. The farm isolates consisted of 512 environmental *Salmonella enterica* isolates, from 30 *Salmonella enterica* serovars, collected from 49 locations, between 2020 and 2021, on and near the onion farms implicated by the epidemiological traceback as well as the nearby irrigation district and public lands.

5.5.2 Phylogenetic analysis of the clinical clade

The SNP matrix was generated by reference-based SNP analysis implemented in the CFSAN SNP Pipeline [113] with default parameters. The reference genome (NCBI SRA accession SRR12199170) used for SNP analysis by this study was one of the clinical isolates from the outbreak. It was selected from the clinical isolates for having one of the highest N50's and lowest number of contigs. The maximum likelihood phylogeny was inferred from the SNP matrix, using GARLI v2.01 [258] under the General Time-Reversible (GTR) model with Gamma distributed rate heterogeneity, estimate invariant sites, 1000 bootstraps, and 2 categories of variable rates. The phylogeny was rooted using NCBI SRA isolate SRR13685683—a *Salmonella* Newport clinical isolate collected in 2018—as the outgroup.

To explore the phylogenetic neighborhood of the clinical clade (Figure 2), all environmental isolates in the NCBI Pathogen Detection database that were within 1,000 core gene alleles of the clinical clade (using the 1,152 gene cgMLST scheme developed for *Salmonella enterica* at the FDA [115]) were recruited for phylogenetic analysis as described in the previous paragraph. The phylogenies were visualized with FigTree (v1.4.4) [259].

5.5.3 Genome assembly, gene prediction and annotation, and pangenome estimation

The short reads were assembled with SPAdes (v3.13.0) [260] using default settings. For quality control, contigs shorter than 500bp or with less than 10X depth of coverage were removed. Genes in the contigs were predicted and annotated with Prokka (v1.14.5) [261]. Further annotations were obtained running eggNOG-Mapper (v2.1.6) [262], using default parameters, on the predicted genes. The predicted genes were clustered with Roary (v3.12.0) [263] to identify the pangenome of the clinical clade using 90% identity for clustering and requiring that 99% of isolates possess a gene to be considered core.

5.5.4 Identification and annotation of plasmids

Platon (v1.6) [264], using default settings, was used to identify and annotate the plasmids in the assemblies. Platon is a tool for the identification of extrachromosomal plasmid contigs in short read draft assemblies. Contigs are characterized by testing for circularization; the detection of incompatibility groups; the detection of rRNA genes; the detection of antimicrobial resistance genes; a homology search against reference plasmid sequences; the detection of oriT sequences; the detection of plasmid replication genes; the detection of mobilization genes; and the detection of conjugation genes [264]. The online COPLA server [244] was used to determine the host range of the plasmids.

The most observed plasmid type was the IncFII(S), annotated as present by Platon in 98.5% of the isolates. The IncFII(S) plasmids, identified by Platon, were BLAST aligned against the pMLST [256] database to identify the sequence profiles. To determine if the IncFII(S) plasmid was present in all clinical isolates, the raw reads of all isolates were mapped to the

IncFII(S) plasmid of isolate SRR12199170 using Bowtie2 [77] with default settings. A custom Python script was used to calculate the breadth of coverage of mapped reads for each isolate. Additionally, to identify all contigs belonging to IncFII(S) in all isolate assemblies, the contigs were BLAST aligned (95% identity) to the SRR12199170 plasmid. For the 24 isolates with no identified IncFII(S) plasmid, 23 had reads mapping, with >92% breadth of coverage, to the complete IncFII(S) plasmid in isolate SRR12199170, and one isolate had 7% breadth of coverage.

5.5.5 Comparison of clinical and farm plasmids and finding the best BLAST hits in NCBI

The clinical and farm isolate plasmids were pairwise BLAST aligned (using 95% identity). The farm plasmids were filtered for those that aligned with at least 90% coverage of a clinical plasmid using a custom Python script. Metadata internal to the FDA was used to determine the GPS coordinates of the sampling locations and the *Salmonella* serovars.

Farm and clinical plasmids that BLAST aligned, were then queried against the NCBI Pathogen Detection and Nucleotide databases using BLAST (≥95% identity). The farm plasmids were filtered for those that aligned with at least 90% coverage of a clinical plasmid using a custom Python script. The results were additionally filtered with a custom Python script for hits that were at least as similar as the farm and clinical plasmids were to each other.

The metadata for the best BLAST hits was either parsed from the GenBank file online, if from NCBI Nucleotide database, or from the metadata provided on the NCBI Pathogen Detection website (https://www.ncbi.nlm.nih.gov/pathogens/). The metadata was used to select for environmental isolates and to filter out clinical isolates so that information about geographic location and the isolation source could be analyzed.

5.5.6 Plasmid phylogenies, mutation graphs, and relatedness

The matching clinical and farm plasmids, as well as the filtered BLAST hits of the clinical plasmids from the NCBI Pathogen Detection and Nucleotide databases, as described earlier, were used to build maximum likelihood phylogenies—this was also done for the IncFII(S) plasmid and its best BLAST hits, although there was no matching farm plasmid.

For the Col440I, ColpVC, and the uncharacterized plasmids, a custom Python script was used to trim the sequences if they were over-circularized and to reorient the sequences to a common start locus. Then whole plasmid alignment was performed with Muscle [265] within the MEGA software (version 11) [266].

For the IncI1-I(Gamma) and IncFII(S) plasmids (Figure 5.4 and Figure 5.6, respectively), recombination, gene gain and loss, and potentially assembly errors made it difficult to perform whole plasmid alignment, so Roary was used to find the core genes (>95% identity, 100% of isolates must have gene to be core), which were then concatenated and aligned by Roary using MAFFT (v7.305b) [267]. There were 41 and 33 core genes for the IncFII(S) and IncI1-I(Gamma) plasmids, respectively. For the IncFII(S) plasmids, only one clinical instance was used (isolate SRR12199170) because all the clinical instances clustered at >99% identity and >95% coverage with MMSeqs2 [268].

Maximum likelihood phylogenies were built from the multiple sequence alignments using Mega with the Tamura-Nei nucleotide substitution model; having invariant sites; using the nearest-neighbor-interchange maximum likelihood heuristic method; using the default neighbor joining method to make the initial tree; the bootstrap method as the test of phylogeny with 1000 bootstraps; using the isolate with the oldest collection date as the root; and collapsing all branches with less than 0.75 bootstrap support. The phylogenies were visualized in FigTree.

In some cases, the phylogenies lacked resolution, i.e., formed large polytomies, and so graphs (referred to here as mutation graphs) were built to visualize the mutational similarities and differences between the plasmid sequences. Mutation graphs were constructed for the Col440I (Figure 5.6), ColpVC (Figure 5.7), and IncFII(S) (Figure 5.8) plasmids by identifying all variations in the multiple sequence alignments between the sequences and the respective clinical plasmids using a custom Python script. The mutation graphs visualize the mutational differences between plasmids and are not intended as phylogenetic analyses [239].

Snippy [269] was used to determine the number of SNP differences between the uncharacterized plasmids. A custom Python script was used to measure the total number of pairwise nucleotide differences for the sequences in the concatenated core gene alignment of the IncI1-I(Gamma) plasmid. In this case, differences were measured per column of the multiple sequence alignment for every two sequence combinations. Here a nucleotide difference could be a mismatch between different nucleotides or a nucleotide and a gap—caused by an insertion or deletion.

5.5.7 Statistical analysis for assessing frequency of observing highly similar plasmids in environmental *Salmonella enterica*

The metadata for *Salmonella enterica* was downloaded from the NCBI Pathogen Detection ftp site (https://ftp.ncbi.nlm.nih.gov/pathogen/Results/*Salmonella*/PDG000000002.2417/Metadata/PDG000000002.2417.metadata.tsv) on March 28, 2022. A custom Python script was used to randomly select 516 environmental isolates. All the clinical ColpVC, Col440I, and IncI1-I(Gamma) plasmids were BLAST aligned to the 516 assemblies of the randomly selected environmental isolates. A custom Python script was used to parse the BLAST results. To calculate the percent identity of all local alignments to a clinical plasmid, the clinical plasmid was initially represented as an array of zeros. Every position within the clinical plasmid was updated with the maximum percent identity from all local alignments. If the average percent identity of the array representing the alignment to the clinical plasmid was as high as the minimum percent identity for the relevant plasmid type (Table 5.3), it was counted as a highly similar. Random sampling of the environmental isolates and finding highly similar plasmids was repeated 10,000 times to build distributions. The distributions were used to assess the frequency of observing the highly similar plasmids. A custom Python script was used to visualize the results.

6. Conclusion

Metagenomics is an invaluable tool for the culture independent exploration of microbial communities. It has shed light on the immense genetic and functional diversity of microbes on Earth and has many important applications, often at the intersection of microbiomes and the health of humans and ecosystems. We believe the methods we developed will help future studies extract biologically meaningful information from metagenomic datasets and that our benchmarking analyses will help future studies develop improved methods and avoid pitfalls. Our analyses also highlighted many opportunities for future work and detailed challenges that need to be addressed to increase the utility of metagenomics for microbiome research.

A key challenge for metagenomics is documenting and organizing all the sequence data extracted from microbiomes to provide a reference for the research community. Microbial gene catalogs attempt to provide such a reference, but our assessment of the methods used to construct such catalogs and their effectiveness when used for metagenomic analysis revealed many issues. Our analysis of the methods used to create gene catalogs highlighted the need for efficient gene clustering algorithms that account for the differential evolution of genes and that dynamically find biologically meaningful clusters [270]. An important consideration for such algorithms is that microbial genes can have different taxonomic host ranges, such as those carried by and horizontally transferred on plasmids. Further, for gene catalogs to be useful as global resources for metagenomic data analysis, new methods need to be developed for handling metagenomic reads with non-specific mappings to a gene catalog. Several approaches have been developed for RNA-seq analysis that effectively handle multi-mappings in an alignment-free manner [68-70], but it remains to be seen whether these are effective in metagenomic settings or whether the underlying algorithms need to be adapted.

A limitation of gene catalogs is that they do not include all the sequence information contained by metagenomic contigs. For example, gene catalogs do not include intergenic sequences and likely do not contain many intron-carrying eukaryotic genes which are more difficult to identify in metagenomic contigs than bacterial genes [271]. One strategy for more comprehensively documenting the sequence information from a microbiome, that has gained traction since we published our study, is the creation of metagenome-assembled-genome (MAG) catalogs [109]. Future work needs to critically evaluate the methods used to build MAG catalogs, including metagenomic binning—the grouping of metagenomic contigs to individual genomes—which is error prone and an active area of research [272, 273]. Additionally, it is unclear if MAG catalogs can efficiently document the pangenome—the total sequence repertoire—of each species in a microbiome, including all genomic variations and structures, given that the genomes of many microbial species frequently undergo horizontal transfer, recombination, and rearrangement.

Another key bioinformatic challenge for the investigation of microbiomes is the comprehensive taxonomic classification of metagenomes. Currently, this is not possible due to the immense amount of microbial dark matter on Earth and efforts to build classification tools have disproportionately focused on prokaryotes versus eukaryotes [17, 18, 184]. We presented taxaTarget, a computationally efficient, sensitive, and precise supervised learning method that represents an important step forward for the identification of eukaryotic species in metagenomic data. Our analysis highlighted the added value of using data-driven approaches versus arbitrary, user-defined thresholds for classification. We believe taxaTarget will help facilitate our understanding of the eukaryotic component of microbiomes and is currently being applied to understand the composition of toxic algae blooms in the Indian River Lagoon which have led to the extensive loss of seagrasses and the starvation of manatees [274].

Despite the performance gains of taxaTarget compared to existing tools, a persistent source of classification errors was sparse training data. As such, initiatives like the Earth BioGenome Project [275] are imperative for increasing the number of sequenced genomes, transcriptomes, and proteomes for underrepresented phylogroups. As more species become represented in sequence databases, the incorporation of more sophisticated machine learning approaches (e.g., logistic regression, support vector machines) and the use of imputation in taxaTarget should be tested. Additionally, the methods of taxaTarget should be tested when extended to other phylogroups (e.g., bacteria, archaea, viruses) and other sequences (perhaps whole genomes) besides marker genes. Ideally, the methods used to train taxaTarget could identify the region-specific discriminatory power for all the sequences in a database and the inclusion of a larger percentage of genomes might improve the sensitivity of classification. However, the use of sequences other than marker genes might not currently be feasible without the development of additional algorithms for handling scarce data and the resulting classification errors.

A key bioinformatic challenge that underlies many of the issues with taxonomic classification as well as gene and MAG catalogs is our current inability to assemble complete genomes for all organisms in all metagenomes. Often, metagenomic assemblies are highly fragmented, making it challenging to associate the contigs to their respective organisms and taxa to their genomes, genes, and functions. In part, this is due to the impractical cost of sequencing to a sufficient depth to obtain complete coverage of all the genomes in a complex microbial community. This is especially an issue for eukaryotes and was highlighted by a recent study where nearly 300 billion reads were required to recover just 700 eukaryotic MAGs [276]. However, even provided sufficient coverage, it can still be challenging to reconstruct complete genomes, especially when using short reads which cannot span all intra- and inter-genomic repeats. Additionally, current assembly approaches often only reconstruct the most abundant strain in the metagenomes, requiring further analysis to reveal the strain structure hidden by the consensus assembly.

Undoubtedly, long read sequencing will increasingly facilitate the more complete reconstruction of genomes from metagenomes, providing insights into genome structure, gene order, and mobile elements that were often unavailable with short read sequencing. However, as shown in Chapter 3, the sequencing error rate needs to decrease before the reconstructed genomes are accurate enough to be useful for pathogen detection and source tracking investigations where high fidelity is critical. Nonetheless, our analysis indicates that the accuracy of the long read assembly algorithms can be improved with better use of the information in the long reads. Currently, the integration of long and short read sequencing by hybrid assemblers provides the highest-quality reconstruction of pathogen genomes from the quasimetagenomes compared to either platform alone. However, the observed differences between hybrid assembly approaches that start with short reads versus those that start with long reads, suggest that hybrid approaches are currently limited by the weaknesses of the different technologies, but can be improved by effectively integrating the contiguity of long reads and high per-base quality of the short reads.

In addition to the many uncharacterized taxa in metagenomic samples, there are also many uncharacterized mobile genetic elements (e.g., phages, plasmids, transposons) which are a fundamental source of adaptation for many microbes and connect microbes via a network of genetic information exchange [224, 225]. Our analysis of horizontally transferred plasmids in Chapter 5 showed that these elements have potential application for relating lineages to microbiomes and environments, perhaps an important extension of WGS source tracking methods. However, our analysis highlighted the need for more detailed and accurate metadata, more extensive environmental sampling, and a better understanding of plasmid molecular evolution. In part, future work needs to characterize the pangenomes of different plasmid types and the gene exchange networks between plasmids (some studies have already initiated such efforts [221, 277]). Such analyses will improve plasmid typing methods and the assessment of the evolutionary relatedness of plasmid sequences. Future work is also needed to understand the distribution, population dynamics, and transmission dynamics of plasmids in the environment [278]. In a metagenomic context, in can be difficult to assemble low abundance and highly repetitive plasmids. It can also be difficult to associate plasmids reconstructed from metagenomes with their host chromosome; however, tools like Hi-C sequencing, the use of methylation patterns detected in long reads, and binning methods have begun making progress on this issue [279, 280].

In conclusion, with the continued improvement of metagenomic tools and technologies it is likely that it will soon be possible to completely and accurately reconstruct most of the genomes in a metagenomic sample. This will enable researchers to work on even more challenging problems like characterizing all the functions coded in microbial genomes and microbiomes. Such efforts will likely require the integration of metagenomics with other meta-omic approaches like metabolomics, metatranscriptomics, and metaproteomics and will lead to a much deeper understanding of the ecological contributions of microbes to their microbial communities, hosts, and ecosystems.

Bibliography

1. Dodd MS, Papineau D, Grenne T, Slack JF, Rittner M, Pirajno F, O'Neil J, Little CTS: **Evidence for early life in Earth's oldest hydrothermal vent precipitates**. *Nature* 2017, **543**(7643):60-64.
2. DasSarma P, DasSarma S: **Survival of microbes in Earth's stratosphere**. *Current Opinion in Microbiology* 2018, **43**:24-30.
3. Gross M: **Life underground**. *Current Biology* 2021, **31**(9):R415-R417.
4. Prieto-Barajas CM, Valencia-Cantero E, Santoyo G: **Microbial mat ecosystems: Structure types, functional diversity, and biotechnological application**. *Electronic Journal of Biotechnology* 2018, **31**:48-56.
5. Dick GJ: **The microbiomes of deep-sea hydrothermal vents: distributed globally, shaped locally**. *Nature Reviews Microbiology* 2019, **17**(5):271-283.
6. Zhong Z-P, Tian F, Roux S, Gazitúa MC, Solonenko NE, Li Y-F, Davis ME, Van Etten JL, Mosley-Thompson E, Rich VI *et al*: **Glacier ice archives nearly 15,000-year-old microbes and phages**. *Microbiome* 2021, **9**(1):160.
7. Lloyd Karen G, Steen Andrew D, Ladau J, Yin J, Crosby L, Neufeld Josh D: **Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes**. *mSystems*, **3**(5):e00055-00018.
8. Das B, Nair GB: **Homeostasis and dysbiosis of the gut microbiome in health and disease**. *Journal of Biosciences* 2019, **44**(5):117.
9. Trinh P, Zaneveld JR, Safranek S, Rabinowitz PM: **One Health Relationships Between Human, Animal, and Environmental Microbiomes: A Mini-Review**. *Frontiers in Public Health* 2018, **6**:235.
10. Handelsman J: **Metagenomics: application of genomics to uncultured microorganisms**. *Microbiol Mol Biol Rev* 2004, **68**(4):669-685.
11. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database C: **The Sequence Read Archive: explosive growth of sequencing data**. *Nucleic acids research* 2012, **40**(Database issue):D54-D56.
12. Locey Kenneth J, Lennon Jay T: **Scaling laws predict global microbial diversity**. *Proceedings of the National Academy of Sciences* 2016, **113**(21):5970-5975.
13. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C: **The Third Revolution in Sequencing Technology**. *Trends in Genetics* 2018, **34**(9):666-681.
14. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, Rasmussen M, Zervas A, Hansen LH: **GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms**. *GigaScience* 2020, **9**(2):giaa008.
15. Delahaye C, Nicolas J: **Sequencing DNA with nanopores: Troubles and biases**. *PLOS ONE* 2021, **16**(10):e0257521.
16. New FN, Brito IL: **What Is Metagenomics Teaching Us, and What Is Missed?** *Annual Review of Microbiology* 2020, **74**(1):117-135.
17. Jiao J-Y, Liu L, Hua Z-S, Fang B-Z, Zhou E-M, Salam N, Hedlund BP, Li W-J: **Microbial dark matter coming to light: challenges and opportunities**. *National Science Review* 2021, **8**(3):nwaa280.
18. Zha Y, Chong H, Yang P, Ning K: **Microbial dark matter: from discovery to applications**. *Genomics, Proteomics & Bioinformatics* 2022.

19. Ottesen A, Ramachandran P, Chen Y, Brown E, Reed E, Strain E: **Quasimetagenomic source tracking of Listeria monocytogenes from naturally contaminated ice cream**. *BMC Infectious Diseases* 2020, **20**(1):83.
20. Hyeon J-Y, Li S, Mann DA, Zhang S, Li Z, Chen Y, Deng X: **Quasimetagenomics-Based and Real-Time-Sequencing-Aided Detection and Subtyping of Salmonella enterica from Food Samples**. *Appl Environ Microbiol* 2018, **84**(4).
21. Townsend A, Li S, Mann DA, Deng X: **A quasimetagenomics method for concerted detection and subtyping of Salmonella enterica and E. coli O157:H7 from romaine lettuce**. *Food Microbiology* 2020, **92**:103575.
22. Ghurye JS, Cepeda-Espinoza V, Pop M: **Metagenomic Assembly: Overview, Challenges and Applications**. *Yale J Biol Med* 2016, **89**(3):353-362.
23. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS: **Long reads: their purpose and place**. *Human Molecular Genetics* 2018, **27**(R2):R234-R241.
24. Pearman WS, Freed NE, Silander OK: **Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads**. *BMC Bioinformatics* 2020, **21**(1):220.
25. Nayfach S, Pollard KS: **Toward Accurate and Quantitative Comparative Metagenomics**. *Cell* 2016, **166**(5):1103-1116.
26. Agrawal K, Verma P: **Metagenomics: a possible solution for uncovering the "mystery box" of microbial communities involved in the treatment of wastewater**. In: *Wastewater Treatment.* Elsevier; 2021: 41-53.
27. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters**. *Bioinformatics* 2007, **23**(10):1282-1288.
28. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic acids research* 2016, **44**(D1):D67-D72.
29. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W: **The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families**. *PLoS Biol* 2007, **5**(3):e16.
30. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D *et al*: **A metagenome-wide association study of gut microbiota in type 2 diabetes**. *Nature* 2012, **490**(7418):55-60.
31. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**(13):1658-1659.
32. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T *et al*: **A human gut microbial gene catalogue established by metagenomic sequencing**. *Nature* 2010, **464**(7285):59-65.
33. Padial JM, De la Riva I: **A paradigm shift in our view of species drives current trends in biological classification**. *Biological Reviews* 2021, **96**(2):731-751.
34. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S: **High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries**. *Nature Communications* 2018, **9**(1):5114.
35. Konstantinidis KT, Ramette A, Tiedje JM: **The bacterial species definition in the genomic era**. *Philos Trans R Soc Lond B Biol Sci* 2006, **361**(1475):1929-1940.
36. Tümmler B: **Molecular epidemiology in current times**. *Environmental Microbiology* 2020, **22**(12):4909-4918.

37. Goering RV: **Pulsed field gel electrophoresis: A review of application and interpretation in the molecular epidemiology of infectious disease**. *Infection, Genetics and Evolution* 2010, **10**(7):866-875.
38. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, Timme R: **Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database**. *Journal of Clinical Microbiology* 2016, **54**(8):1975-1983.
39. Commichaux S, Javkar K, Ramachandran P, Nagarajan N, Bertrand D, Chen Y, Reed E, Gonzalez-Escalona N, Strain E, Rand H *et al*: **Evaluating the accuracy of Listeria monocytogenes assemblies from quasimetagenomic samples using long and short reads**. *BMC Genomics* 2021, **22**(1):389.
40. Patro Jennifer N, Ramachandran P, Barnaba T, Mammel Mark K, Lewis Jada L, Elkins Christopher A, Suen G: **Culture-Independent Metagenomic Surveillance of Commercially Available Probiotics with High-Throughput Next-Generation Sequencing**. *mSphere*, **1**(2):e00057-00016.
41. Commichaux S, Shah N, Ghurye J, Stoppel A, Goodheart JA, Luque GG, Cummings MP, Pop M: **A critical assessment of gene catalogs for metagenomic analysis**. *Bioinformatics* 2021, **37**(18):2848-2857.
42. Rho M, Tang H, Ye Y: **FragGeneScan: predicting genes in short and error-prone reads**. *Nucleic Acids Res* 2010, **38**(20):e191.
43. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W *et al*: **The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families**. *PLoS Biol* 2007, **5**(3):e16.
44. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E *et al*: **Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes**. *Nature biotechnology* 2014, **32**(8):822-828.
45. Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, Li X, Long H, Zhang J, Zhang D *et al*: **A catalog of the mouse gut metagenome**. *Nat Biotechnol* 2015, **33**(10):1103-1108.
46. Pan H, Guo R, Zhu J, Wang Q, Ju Y, Xie Y, Zheng Y, Wang Z, Li T, Liu Z *et al*: **A gene catalogue of the Sprague-Dawley rat gut metagenome**. *Gigascience* 2018, **7**(5).
47. Xiao L, Estellé J, Kiilerich P, Ramayo-Caldas Y, Xia Z, Feng Q, Pedersen AØ, Kjeldsen NJ, Maguin E, Doré J *et al*: **P1016 The pig's other genome: A reference gene catalog of the gut microbiome as a new resource for deep studies of the interplay between the host and its microbiome**. *Journal of Animal Science* 2016, **94**(suppl_4):22-22.
48. Xiao L, Estellé J, Kiilerich P, Ramayo-Caldas Y, Xia Z, Feng Q, Liang S, Pedersen AØ, Kjeldsen NJ, Liu C *et al*: **A reference gene catalogue of the pig gut microbiome**. *Nature Microbiology* 2016, **1**(12):16161.
49. Coelho LP, Kultima JR, Costea PI, Fournier C, Pan Y, Czarnecki-Maulden G, Hayward MR, Forslund SK, Schmidt TSB, Descombes P *et al*: **Similarity of the dog and human gut microbiomes in gene content and response to diet**. *Microbiome* 2018, **6**(1):72.
50. Li J, Zhong H, Ramayo-Caldas Y, Terrapon N, Lombard V, Potocki-Veronese G, Estellé J, Popova M, Yang Z, Zhang H *et al*: **A catalog of microbial genes from the bovine rumen unveils a specialized and diverse biomass-degrading environment**. *GigaScience* 2020, **9**(6).

51. Li X, Liang S, Xia Z, Qu J, Liu H, Liu C, Yang H, Wang J, Madsen L, Hou Y *et al*: **Establishment of a Macaca fascicularis gut microbiome gene catalog and comparison with the human, pig, and mouse gut microbiomes**. *Gigascience* 2018, **7**(9).
52. Huang P, Zhang Y, Xiao K, Jiang F, Wang H, Tang D, Liu D, Liu B, Liu Y, He X *et al*: **The chicken gut metagenome and the modulatory effects of plant-derived benzylisoquinoline alkaloids**. *Microbiome* 2018, **6**(1):211.
53. Mittal P, Saxena R, Gupta A, Mahajan S, Sharma VK: **The Gene Catalog and Comparative Analysis of Gut Microbiome of Big Cats Provide New Insights on Panthera Species**. *Frontiers in Microbiology* 2020, **11**:1012.
54. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A *et al*: **Structure and function of the global ocean microbiome**. *Science* 2015, **348**(6237):1261359.
55. Lou J, Liu M, Gu J, Liu Q, Zhao L, Ma Y, Wei D: **Metagenomic sequencing reveals microbial gene catalogue of phosphinothricin-utilized soils in South China**. *Gene* 2019, **711**:143942.
56. Ma B, France MT, Crabtree J, Holm JB, Humphrys MS, Brotman RM, Ravel J: **A comprehensive non-redundant gene catalog reveals extensive within-community intraspecies diversity in the human vagina**. *Nature Communications* 2020, **11**(1):940.
57. Dai W, Wang H, Zhou Q, Li D, Feng X, Yang Z, Wang W, Qiu C, Lu Z, Xu X *et al*: **An integrated respiratory microbial gene catalogue to better understand the microbial aetiology of Mycoplasma pneumoniae pneumonia**. *GigaScience* 2019, **8**(8).
58. Wang J, Jia H: **Metagenome-wide association studies: fine-mining the microbiome**. *Nat Rev Microbiol* 2016, **14**(8):508-522.
59. Plaza Oñate F, Le Chatelier E, Almeida M, Cervino ACL, Gauthier F, Magoulès F, Ehrlich SD, Pichaud M: **MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data**. *Bioinformatics* 2018:bty830-bty830.
60. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T *et al*: **An integrated catalog of reference genes in the human gut microbiome**. *Nature biotechnology* 2014, **32**(8):834.
61. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution**. *Nucleic Acids Research* 2000, **28**(1):33-36.
62. Sonnhammer ELL, Eddy SR, Durbin R: **Pfam: A comprehensive database of protein domain families based on seed alignments**. *Proteins: Structure, Function, and Bioinformatics* 1997, **28**(3):405-420.
63. Lan R, Reeves PR: **When does a clone deserve a name? A perspective on bacterial species based on population genetics**. *Trends in Microbiology* 2001, **9**(9):419-424.
64. Ghodsi M, Liu B, Pop M: **DNACLUST: accurate and efficient clustering of phylogenetic marker genes**. *BMC Bioinformatics* 2011, **12**(1):271.
65. Edgar RC: **Search and clustering orders of magnitude faster than BLAST**. *Bioinformatics* 2010, **26**(19):2460-2461.
66. Nguyen NP, Warnow T, Pop M, White B: **A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity**. *NPJ Biofilms Microbiomes* 2016, **2**:16004.

67. Shah N, Altschul SF, Pop M: **Outlier detection in BLAST hits**. *Algorithms Mol Biol* 2018, **13**:7.
68. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: **Salmon provides fast and bias-aware quantification of transcript expression**. *Nature Methods* 2017, **14**:417.
69. Patro R, Mount SM, Kingsford C: **Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms**. *Nature biotechnology* 2014, **32**(5):462-464.
70. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq quantification**. *Nature biotechnology* 2016, **34**:525.
71. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics* 2011, **12**(1):323.
72. Routy B, Le Chatelier E, Derosa L, Duong CPM, Alou MT, Daillère R, Fluckiger A, Messaoudene M, Rauber C, Roberti MP *et al*: **Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors**. *Science* 2017, **359**(6371):91-97.
73. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, Costea PI, Godneva A, Kalka IN, Bar N *et al*: **Environment dominates over host genetics in shaping human gut microbiota**. *Nature* 2018, **555**(7695):210-215.
74. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M: **Personalized nutrition by prediction of glycemic responses**. *Cell* 2015, **163**(5):1079-1094.
75. Meziti A, Tsementzi D, Ar Kormas K, Karayanni H, Konstantinidis KT: **Anthropogenic effects on bacterial diversity and function along a river-to-estuary gradient in Northwest Greece revealed by metagenomics**. *Environ Microbiol* 2016, **18**(12):4640-4652.
76. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nat Methods* 2015, **12**(1):59-60.
77. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**(4):357-359.
78. Nyholm O, Lienemann T, Halkilahti J, Mero S, Rimhanen-Finne R, Lehtinen V, Salmenlinna S, Siitonen A: **Characterization of Shigella sonnei Isolate Carrying Shiga Toxin 2-Producing Gene**. *Emerg Infect Dis* 2015, **21**(5):891-892.
79. Lamba K, Nelson JA, Kimura AC, Poe A, Collins J, Kao AS, Cruz L, Inami G, Vaishampayan J, Garza A *et al*: **Shiga Toxin 1-Producing Shigella sonnei Infections, California, United States, 2014-2015**. *Emerg Infect Dis* 2016, **22**(4):679-686.
80. Mattock E, Blocker AJ: **How Do the Virulence Factors of Shigella Work Together to Cause Disease?** *Front Cell Infect Microbiol* 2017, **7**.
81. Juhas M: **Horizontal gene transfer in human pathogens**. *Crit Rev Microbiol* 2013, **41**(1):101-108.
82. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
83. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.
84. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *Journal of Molecular Biology* 1990, **215**(3):403-410.

85. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 2010, **26**(4):493-500.
86. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H *et al*: **Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life**. *Cell Host Microbe* 2015, **17**(5):690-703.
87. Guo Z, Zhang J, Wang Z, Ang KY, Huang S, Hou Q, Su X, Qiao J, Zheng Y, Wang L *et al*: **Intestinal Microbiota Distinguish Gout Patients from Healthy Humans**. *Sci Rep* 2016, **6**:20602.
88. Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, Prifti E, Vieira-Silva S, Gudmundsdottir V, Pedersen HK *et al*: **Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota**. *Nature* 2015, **528**(7581):262-266.
89. Dhakan DB, Maji A, Sharma AK, Saxena R, Pulikkan J, Grace T, Gomez A, Scaria J, Amato KR, Sharma VK: **The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches**. *Gigascience* 2019, **8**(3).
90. Xie H, Guo R, Zhong H, Feng Q, Lan Z, Qin B, Ward KJ, Jackson MA, Xia Y, Chen X *et al*: **Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome**. *Cell Syst* 2016, **3**(6):572-584.e573.
91. Lokmer A, Cian A, Froment A, Gantois N, Viscogliosi E, Chabé M, Ségurel L: **Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels**. *PLoS One* 2019, **14**(2):e0211139.
92. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph**. *Bioinformatics* 2015, **31**(10):1674-1676.
93. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics* 2014, **30**(14):2068-2069.
94. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L *et al*: **Alterations of the human gut microbiome in liver cirrhosis**. *Nature* 2014, **513**(7516):59-64.
95. Vatanen T, Plichta DR, Somani J, Münch PC, Arthur TD, Hall AB, Rudolf S, Oakeley EJ, Ke X, Young RA *et al*: **Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life**. *Nature Microbiology* 2019, **4**(3):470-479.
96. Wang C, Li P, Yan Q, Chen L, Li T, Zhang W, Li H, Chen C, Han X, Zhang S *et al*: **Characterization of the Pig Gut Microbiome and Antibiotic Resistome in Industrialized Feedlots in China**. *mSystems* 2019, **4**(6):e00206-00219.
97. Zheng A, Yi H, Li F, Han L, Yu J, Cheng X, Su H, Hong K, Li J: **Changes in Gut Microbiome Structure and Function of Rats with Isoproterenol-Induced Heart Failure**. *International Heart Journal* 2019, **60**(5):1176-1183.
98. Lesker TR, Durairaj AC, Gálvez EJC, Lagkouvardos I, Baines JF, Clavel T, Sczyrba A, McHardy AC, Strowig T: **An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome**. *Cell Reports* 2020, **30**(9):2909-2922.e2906.
99. Liu G, Wu C, Abrams WR, Li Y: **Structural and Functional Characteristics of the Microbiome in Deep-Dentin Caries**. *Journal of Dental Research* 2020, **99**(6):713-720.

100. Zhou Y, Coventry DR, Vadakattu VSRG, Fuentes D, Merchant A, Kaiser BN, Li J, Wei Y, Liu H, Wang Y *et al*: **The preceding root system drives the composition and function of the rhizosphere microbiome**. *Genome Biology* 2020, **21**:1-19.
101. Ghanbari Maman L, Palizban F, Fallah Atanaki F, Elmi Ghiasi N, Ariaeenejad S, Ghaffari MR, Hosseini Salekdeh G, Kavousi K: **Co-abundance analysis reveals hidden players associated with high methane yield phenotype in sheep rumen microbiome**. *Scientific Reports* 2020, **10**(1):4995.
102. McLaren MR, Willis AD, Callahan BJ: **Consistent and correctable bias in metagenomic sequencing experiments**. *eLife* 2019, **8**:e46923.
103. Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL: **Confronting the catalytic dark matter encoded by sequenced genomes**. *Nucleic Acids Research* 2017, **45**(20):11495-11514.
104. Nasko DJ, Koren S, Phillippy AM, Treangen TJ: **RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification**. *Genome Biology* 2018, **19**(1):165.
105. Navlakha S, White J, Nagarajan N, Pop M, Kingsford C: **Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information**. *J Comput Biol* 2010, **17**(3):503-516.
106. White JR, Navlakha S, Nagarajan N, Ghodsi MR, Kingsford C, Pop M: **Alignment and clustering of phylogenetic markers--implications for microbial diversity studies**. *BMC Bioinformatics* 2010, **11**:152.
107. Hao X, Jiang R, Chen T: **Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering**. *Bioinformatics* 2011, **27**(5):611-618.
108. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP: **DADA2: High-resolution sample inference from Illumina amplicon data**. *Nat Methods* 2016, **13**(7):581-583.
109. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P *et al*: **A unified catalog of 204,938 reference genomes from the human gut microbiome**. *Nature biotechnology* 2020.
110. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, the CDCPTF: **PulseNet: The Molecular Subtyping Network for Foodborne Bacterial Disease Surveillance, United States**. *Emerging Infectious Diseases* 2001, **7**(3):382-389.
111. Centers for Disease C, Prevention: **Establishment of a national surveillance program for antimicrobial resistance in Salmonella**. *MMWR Morb Mortal Wkly Rep* 1996, **45**(5):110-111.
112. Tollefson L: **FDA reveals plans for antimicrobial susceptibility monitoring**. *J Am Vet Med Assoc* 1996, **208**(4):459-460.
113. Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, Strain E: **CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data**. *PeerJ Computer Science* 2015, **1**:e20.
114. Pightling AW, Petronella N, Pagotto F: **The Listeria monocytogenes Core-Genome Sequence Typer (LmCGST): a bioinformatic pipeline for molecular characterization with next-generation sequence data**. *BMC Microbiol* 2015, **15**:224.
115. Pettengill JB, Pightling AW, Baugher JD, Rand H, Strain E: **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* 2016, **11**(11):e0166162.
116. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F, Harmsen D, Mellmann A: **Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of Listeria monocytogenes**. *J Clin Microbiol* 2015, **53**(9):2869-2876.
117. Mellmann A, Bletz S, Böking T, Kipp F, Becker K, Schultes A, Prior K, Harmsen D: **Real-Time Genome Sequencing of Resistant Bacteria Provides Precision Infection Control in an Institutional Setting**. *J Clin Microbiol* 2016, **54**(12):2874-2881.
118. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M: **A genomic overview of the population structure of Salmonella**. *PLoS Genet* 2018, **14**(4):e1007261.
119. Pearce ME, Alikhan N-F, Dallman TJ, Zhou Z, Grant K, Maiden MCJ: **Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak**. *Int J Food Microbiol* 2018, **274**:1-11.
120. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C: **Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data**. *BMC Bioinformatics* 2016, **17**:125.
121. Moss EL, Maghini DG, Bhatt AS: **Complete, closed bacterial genomes from microbiomes using nanopore sequencing**. *Nat Biotechnol* 2020.
122. Wang L, Qu L, Yang L, Wang Y, Zhu H: **NanoReviser: An Error-Correction Tool for Nanopore Sequencing Based on a Deep Learning Algorithm**. *Front Genet* 2020, **11**:900.
123. Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H: **Benchmarking of long-read correction methods**. *NAR Genom Bioinform* 2020, **2**(2):lqaa037.
124. Lemon JK, Khil PP, Frank KM, Dekker JP: **Rapid Nanopore Sequencing of Plasmids and Resistance Gene Detection in Clinical Isolates**. *J Clin Microbiol* 2017, **55**(12):3530-3543.
125. Chen Z, Kuang D, Xu X, González-Escalona N, Erickson DL, Brown E, Meng J: **Genomic analyses of multidrug-resistant Salmonella Indiana, Typhimurium, and Enteritidis isolates using MinION and MiSeq sequencing technologies**. *PLoS One* 2020, **15**(7):e0235641.
126. González-Escalona N, Allard MA, Brown EW, Sharma S, Hoffmann M: **Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing Escherichia coli**. *PLoS One* 2019, **14**(7):e0220494.
127. United States F, Drug A: **Bacteriological Analytical Manual**: Aoac International; 1995.
128. Ottesen A, Ramachandran P, Reed E, White JR, Hasan N, Subramanian P, Ryan G, Jarvis K, Grim C, Daquiqan N *et al*: **Enrichment dynamics of Listeria monocytogenes and the associated microbiome from naturally contaminated ice cream linked to a listeriosis outbreak**. *BMC Microbiol* 2016, **16**(1):275.
129. Loit K, Adamson K, Bahram M, Puusepp R, Anslan S, Kiiker R, Drenkhan R, Tedersoo L: **Relative Performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) Third-Generation Sequencing Instruments in Identification of Agricultural and Forest Fungal Pathogens**. *Appl Environ Microbiol* 2019, **85**(21).

130. Petersen LM, Martin IW, Moschetti WE, Kershaw CM, Tsongalis GJ: **Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing**. *J Clin Microbiol* 2019, **58**(1).
131. Hamner S, Brown BL, Hasan NA, Franklin MJ, Doyle J, Eggers MJ, Colwell RR, Ford TE: **Metagenomic Profiling of Microbial Pathogens in the Little Bighorn River, Montana**. *Int J Environ Res Public Health* 2019, **16**(7).
132. Latorre-Pérez A, Villalba-Bermell P, Pascual J, Vilanova C: **Assembly methods for nanopore-based metagenomic sequencing: a comparative study**. *Sci Rep* 2020, **10**(1):13588.
133. Pettengill JB, McAvoy E, White JR, Allard M, Brown E, Ottesen A: **Using metagenomic analyses to estimate the consequences of enrichment bias for pathogen detection**. *BMC Res Notes* 2012, **5**:378.
134. Ottesen AR, Gonzalez A, Bell R, Arce C, Rideout S, Allard M, Evans P, Strain E, Musser S, Knight R *et al*: **Co-enriching microflora associated with culture based methods to detect Salmonella from tomato phyllosphere**. *PLoS One* 2013, **8**(9):e73079.
135. Lusk TS, Ottesen AR, White JR, Allard MW, Brown EW, Kase JA: **Characterization of microflora in Latin-style cheeses by next-generation sequencing technology**. *BMC Microbiol* 2012, **12**:254.
136. Jarvis KG, White JR, Grim CJ, Ewing L, Ottesen AR, Beaubrun JJ-G, Pettengill JB, Brown E, Hanes DE: **Cilantro microbiome before and after nonselective pre-enrichment for Salmonella using 16S rRNA and metagenomic sequencing**. *BMC Microbiol* 2015, **15**:160.
137. Lusk Pfefer T, Ramachandran P, Reed E, Kase JA, Ottesen A: **Metagenomic Description of Preenrichment and Postenrichment of Recalled Chapati Atta Flour Using a Shotgun Sequencing Approach**. *Genome Announc* 2018, **6**(21).
138. Ramachandran P, Reed E, Ottesen A: **Exploring the Microbiome of Callinectes sapidus (Maryland Blue Crab)**. *Genome Announc* 2018, **6**(22).
139. Ossai S, Ramachandran P, Ottesen A, Reed E, DePaola A, Parveen S: **Microbiomes of American Oysters () Harvested from Two Sites in the Chesapeake Bay**. *Genome Announc* 2017, **5**(30).
140. Ramachandran P, Reed E, Commichaux S, Strain E, Depaola A, Rikard S, Ottesen A: **Characterization of the Microbiota of Oyster Larvae (Crassostrea virginica) and Tank Water from an Aquaculture System with High and Low Larval Survival Rates**. *Genome Announc* 2018, **6**(25).
141. Chen Y, Luo Y, Curry P, Timme R, Melka D, Doyle M, Parish M, Hammack TS, Allard MW, Brown EW *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* 2017, **12**(2):e0171389.
142. Roberts RJ, Carneiro MO, Schatz MC: **The advantages of SMRT sequencing**. *Genome Biol* 2013, **14**(7):405.
143. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM: **Mash: fast genome and metagenome distance estimation using MinHash**. *Genome Biol* 2016, **17**(1):132.
144. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, Dvornicic M, Soldo JP, Koh JY, Tong C *et al*: **Hybrid metagenomic assembly enables high-resolution**

analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol* 2019, **37**(8):937-944.
145. **[No title]**.
146. **Basecalling using Guppy**.
147. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes**. *Nucleic Acids Res* 2007, **35**(9):3100-3108.
148. tseemann: **tseemann/snippy**.
149. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
150. Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics* 2018, **34**(18):3094-3100.
151. Ghurye J, Treangen T, Fedarko M, Hervey WJt, Pop M: **MetaCarvel: linking assembly graph motifs to biological variants**. *Genome Biol* 2019, **20**(1):174.
152. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation**. *Genome Research* 2017, **27**(5):722-736.
153. Ruan J, Li H: **Fast and accurate long-read assembly with wtdbg2**. *Nat Methods* 2020, **17**(2):155-158.
154. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL *et al*: **metaFlye: scalable long-read metagenome assembly using repeat graphs**. *Nat Methods* 2020, **17**(11):1103-1110.
155. Vaser R, Sović I, Nagarajan N, Šikić M: **Fast and accurate de novo genome assembly from long uncorrected reads**. *Genome Research* 2017, **27**(5):737-746.
156. Antipov D, Korobeynikov A, McLean JS, Pevzner PA: **hybridSPAdes: an algorithm for hybrid assembly of short and long reads**. *Bioinformatics* 2016, **32**(7):1009-1015.
157. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK *et al*: **Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement**. *PLoS One* 2014, **9**(11):e112963.
158. Warren RL, Coombe L, Mohamadi H, Zhang J, Jaquish B, Isabel N, Jones SJM, Bousquet J, Bohlmann J, Birol I: **ntEdit: scalable genome sequence polishing**. *Bioinformatics* 2019, **35**(21):4430-4432.
159. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: quality assessment tool for genome assemblies**. *Bioinformatics* 2013, **29**(8):1072-1075.
160. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Res* 2008, **18**(11):1851-1858.
161. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments**. *Genome Biology* 2014, **15**(3):R46.
162. Commichaux S, Javkar K, Muralidharan HS, Ramachandran P, Ottesen A, Rand H, Pop M: **TaxaTarget: Fast, Sensitive, and Precise Classification of Microeukaryotes in Metagenomic Data**. 2021.
163. Hawksworth DL, Lücking R: **Fungal Diversity Revisited: 2.2 to 3.8 Million Species**. *Microbiol Spectr* 2017, **5**(4).

164. Larsen BB, Miller EC, Rhodes MK, Wiens JJ: **Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life**. *The Quarterly Review of Biology* 2017, **92**(3):229-265.
165. Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, Bowser SS, Cepicka I, Decelle J, Dunthorn M *et al*: **CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms**. *PLoS Biology* 2012, **10**(11):e1001419.
166. Burki F, Roger AJ, Brown MW, Simpson AGB: **The New Tree of Eukaryotes**. *Trends in Ecology & Evolution* 2020, **35**(1):43-55.
167. Rousseaux C, Gregg W: **Interannual Variation in Phytoplankton Primary Production at A Global Scale**. *Remote Sensing* 2013, **6**(1):1-19.
168. Wong Z-W, Faulder K, Robinson JL: **Does Dientamoeba fragilis cause diarrhea? A systematic review**. *Parasitol Res* 2018, **117**(4):971-980.
169. Deng L, Wojciech L, Gascoigne NRJ, Peng G, Tan KSW: **New insights into the interactions between Blastocystis, the gut microbiota, and host immunity**. *PLoS Pathog* 2021, **17**(2):e1009253.
170. Hirt RP: **Mucosal microbial parasites/symbionts in health and disease: an integrative overview**. *Parasitology* 2019, **146**(9):1109-1115.
171. Campo J, Bass D, Keeling PJ: **The eukaryome: Diversity and role of microeukaryotic organisms associated with animal hosts**. *Functional Ecology* 2019.
172. Parfrey LW, Walters WA, Lauber CL, Clemente JC, Berg-Lyons D, Teiling C, Kodira C, Mohiuddin M, Brunelle J, Driscoll M *et al*: **Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity**. *Front Microbiol* 2014, **5**:298.
173. Even G, Lokmer A, Rodrigues J, Audebert C, Viscogliosi E, Ségurel L, Chabé M: **Changes in the Human Gut Microbiota Associated With Colonization by sp. and spp. in Non-Industrialized Populations**. *Front Cell Infect Microbiol* 2021, **11**:533528.
174. New FN, Brito IL: **What Is Metagenomics Teaching Us, and What Is Missed?** *Annu Rev Microbiol* 2020, **74**:117-135.
175. Laforest-Lapointe I, Arrieta M-C: **Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies**. *mSystems* 2018, **3**(2).
176. Hooks KB, O'Malley MA: **Contrasting Strategies: Human Eukaryotic Versus Bacterial Microbiome Research**. *J Eukaryot Microbiol* 2019.
177. Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB: **Protists are microbes too: a perspective**. *The ISME Journal* 2009, **3**(1):4-12.
178. Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK: **Sequencing our way towards understanding global eukaryotic biodiversity**. *Trends Ecol Evol* 2012, **27**(4):233-243.
179. Jacquiod S, Stenbæk J, Santos SS, Winding A, Sørensen SJ, Priemé A: **Metagenomes provide valuable comparative information on soil microeukaryotes**. *Research in Microbiology* 2016, **167**(5):436-450.
180. Saary P, Mitchell AL, Finn RD: **Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC**. *Genome Biol* 2020, **21**(1):244.
181. West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF: **Genome-reconstruction for eukaryotes from complex natural microbial communities**. *Genome Res* 2018, **28**(4):569-580.

182. Lu J, Salzberg SL: **Removing contaminants from databases of draft genomes**. *PLoS Comput Biol* 2018, **14**(6):e1006277.
183. Steinegger M, Salzberg SL: **Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank**. *Genome Biol* 2020, **21**(1):115.
184. Lind AL, Pollard KS: **Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing**. *Microbiome* 2021, **9**(1):58.
185. Nousias O, Montesanto F: **Metagenomic profiling of host-associated bacteria from 8 datasets of the red alga Porphyra purpurea with MetaPhlAn3**. *Mar Genomics* 2021:100866.
186. Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH: **METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data**. *Mol Ecol Resour* 2015, **15**(6):1403-1414.
187. Salazar G, Ruscheweyh H-J, Hildebrand F, Acinas SG, Sunagawa S: **mTAGs: taxonomic profiling using degenerate consensus reference sequences of ribosomal RNA genes**. *Bioinformatics* 2022, **38**(1):270-272.
188. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.
189. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ: **The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages**. *Proc Natl Acad Sci U S A* 2009, **106**(18):7273-7280.
190. Marcelino VR, Clausen PTLC, Buchmann JP, Wille M, Iredell JR, Meyer W, Lund O, Sorrell TC, Holmes EC: **CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data**. *Genome Biol* 2020, **21**(1):103.
191. Shah N, Meisel JS, Pop M: **Embracing Ambiguity in the Taxonomic Classification of Microbiome Sequencing Data**. *Front Genet* 2019, **10**:1022.
192. Liu B, Gibbons T, Ghodsi M, Pop M: **MetaPhyler: Taxonomic profiling for metagenomic sequences**. *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2010.
193. Orellana LH, Rodriguez-R LM, Konstantinidis KT: **ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores**. *Nucleic Acids Res* 2017, **45**(3):e14.
194. Menzel P, Ng KL, Krogh A: **Fast and sensitive taxonomic classification for metagenomics with Kaiju**. *Nat Commun* 2016, **7**:11257.
195. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nature Methods* 2015, **12**(1):59-60.
196. Consortium TU, The UniProt C: **The Universal Protein Resource (UniProt)**. *Nucleic Acids Research* 2007, **36**(Database):D190-D195.
197. Huang W, Li L, Myers JR, Marth GT: **ART: a next-generation sequencing read simulator**. *Bioinformatics* 2012, **28**(4):593-594.
198. Nash AK, Auchtung TA, Wong MC, Smith DP, Gesell JR, Ross MC, Stewart CJ, Metcalf GA, Muzny DM, Gibbs RA *et al*: **The gut mycobiome of the Human Microbiome Project healthy cohort**. *Microbiome* 2017, **5**(1):153.
199. Nieves-Ramírez ME, Partida-Rodríguez O, Laforest-Lapointe I, Reynolds LA, Brown EM, Valdez-Salazar A, Morán-Silva P, Rojas-Velázquez L, Morien E, Parfrey LW *et al*:

Asymptomatic Intestinal Colonization with Protist Blastocystis Is Strongly Associated with Distinct Microbiome Ecological Patterns. *mSystems* 2018, **3**(3).

200. Iyer LR, Verma AK, Paul J, Bhattacharya A: **Phagocytosis of Gut Bacteria by Entamoeba histolytica**. *Frontiers in Cellular and Infection Microbiology* 2019, **9**.
201. Spatz M, Richard ML: **Overview of the Potential Role of Malassezia in Gut Health and Disease**. *Frontiers in Cellular and Infection Microbiology* 2020, **10**.
202. Hallen-Adams HE, Suhr MJ: **Fungi in the healthy human gastrointestinal tract**. *Virulence* 2017, **8**(3):352-358.
203. Campa-Córdova AI, Luna-González A, Ascencio F, Cortés-Jacinto E, Cáceres-Martínez CJ: **Effects of chloramphenicol, erythromycin, and furazolidone on growth of Isochrysis galbana and Chaetoceros gracilis**. *Aquaculture* 2006, **260**(1-4):145-150.
204. Dimidi E, Cox SR, Rossi M, Whelan K: **Fermented Foods: Definitions and Characteristics, Impact on the Gut Microbiota and Effects on Gastrointestinal Health and Disease**. *Nutrients* 2019, **11**(8).
205. Afrin T, Murase K, Kounosu A, Hunt VL, Bligh M, Maeda Y, Hino A, Maruyama H, Tsai IJ, Kikuchi T: **Sequential Changes in the Host Gut Microbiota During Infection With the Intestinal Parasitic Nematode Strongyloides venezuelensis**. *Frontiers in Cellular and Infection Microbiology* 2019, **9**.
206. Santiago-Rodriguez TM, Fornaciari G, Luciani S, Dowd SE, Toranzos GA, Marota I, Cano RJ: **Taxonomic and predicted metabolic profiles of the human gut microbiome in pre-Columbian mummies**. *FEMS Microbiol Ecol* 2016, **92**(11).
207. Gouba N, Raoult D, Drancourt M: **Eukaryote culturomics of the gut reveals new species**. *PLoS One* 2014, **9**(9):e106994.
208. Berrilli F, Di Cave D, Cavallero S, D'Amelio S: **Interactions between parasites and microbial communities in the human gut**. *Frontiers in Cellular and Infection Microbiology* 2012, **2**.
209. Zajc J, Gunde-Cimerman N: **The Genus From Contamination of Food to Health Threat**. *Microorganisms* 2018, **6**(2).
210. Wijayalath W, Majji S, Kleschenko Y, Pow-Sang L, Brumeanu TD, Villasante EF, Vasta GR, Fernández-Robledo J-A, Casares S: **Humanized HLA-DR4 Mice Fed with the Protozoan Pathogen of Oysters Perkinsus Marinus (Dermo) Do Not Develop Noticeable Pathology but Elicit Systemic Immunity**. *PLoS ONE* 2014, **9**(1):e87435.
211. Dacks J: **Faculty Opinions recommendation of On the reversibility of parasitism: adaptation to a free-living lifestyle via gene acquisitions in the diplomonad Trepomonas sp. PC1**. *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature* 2017.
212. Bucková B, Hurníková Z, Lauková A, Revajová V, Dvorožňáková E: **The anti-parasitic effect of probiotic bacteria via limiting the fecundity of Trichinella spiralis female adults**. *Helminthologia* 2018, **55**(2):102-111.
213. Sibbald SJ, Archibald JM: **More protist genomes needed**. *Nat Ecol Evol* 2017, **1**(5):145.
214. Kang H: **The prevention and handling of the missing data**. *Korean Journal of Anesthesiology* 2013, **64**(5):402.
215. Deng C, Daley T, Calabrese P, Ren J, Smith AD: **Predicting the Number of Bases to Attain Sufficient Coverage in High-Throughput Sequencing Experiments**. *Journal of Computational Biology* 2019, **27**(7):1130-1143.

216. Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, Jones TF, Fazil A, Hoekstra RM, for the International Collaboration on Enteric Disease "Burden of Illness" S: **The Global Burden of Nontyphoidal Salmonella Gastroenteritis**. *Clinical Infectious Diseases* 2010, **50**(6):882-889.
217. Cao G, Meng J, Strain E, Stones R, Pettengill J, Zhao S, McDermott P, Brown E, Allard M: **Phylogenetics and Differentiation of Salmonella Newport Lineages by Whole Genome Sequencing**. *PLOS ONE* 2013, **8**(2):e55687.
218. Pan H, Paudyal N, Li X, Fang W, Yue M: **Multiple Food-Animal-Borne Route in Transmission of Antibiotic-Resistant Salmonella Newport to Humans**. *Frontiers in Microbiology* 2018, **9**:23.
219. You Y, Rankin Shelley C, Aceto Helen W, Benson Charles E, Toth John D, Dou Z: **Survival of Salmonella enterica Serovar Newport in Manure and Manure-Amended Soils**. *Applied and Environmental Microbiology* 2006, **72**(9):5777-5783.
220. **Outbreak of Salmonella Newport Infections Linked to Onions** [https://www.cdc.gov/salmonella/newport-07-20/index.html]
221. Acman M, van Dorp L, Santini JM, Balloux F: **Large-scale network analysis captures biological features of bacterial plasmids**. *Nat Commun* 2020, **11**(1):2452-2452.
222. **Factors Potentially Contributing to the Contamination of Red Onions Implicated in the Summer 2020 Outbreak of Salmonella Newport** [https://www.fda.gov/food/outbreaks-foodborne-illness/factors-potentially-contributing-contamination-red-onions-implicated-summer-2020-outbreak-salmonella]
223. Blanc DS, Magalhães B, Koenig I, Senn L, Grandbastien B: **Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumericsTM) Versus SNP Variant Calling for Epidemiological Investigation of Pseudomonas aeruginosa**. *Frontiers in Microbiology* 2020, **11**:1729.
224. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation**. *Nature* 2000, **405**(6784):299-304.
225. Arnold BJ, Huang IT, Hanage WP: **Horizontal gene transfer and adaptive evolution in bacteria**. *Nature Reviews Microbiology* 2021.
226. Sevillya G, Adato O, Snir S: **Detecting horizontal gene transfer: a probabilistic approach**. *BMC Genomics* 2020, **21**(1):106.
227. Pinilla-Redondo R, Cyriaque V, Jacquiod S, Sørensen SJ, Riber L: **Monitoring plasmid-mediated horizontal gene transfer in microbiomes: recent advances and future perspectives**. *Plasmid* 2018, **99**:56-67.
228. Hall JPJ, Brockhurst MA, Dytham C, Harrison E: **The evolution of plasmid stability: Are infectious transmission and compensatory evolution competing evolutionary trajectories?** *Plasmid* 2017, **91**:90-95.
229. McMillan EA, Jackson CR, Frye JG: **Transferable Plasmids of Salmonella enterica Associated With Antibiotic Resistance Genes**. *Frontiers in Microbiology* 2020, **11**:2497.
230. Cao G, Allard M, Hoffmann M, Muruvanda T, Luo Y, Payne J, Meng K, Zhao S, McDermott P, Brown E *et al*: **Sequence Analysis of IncA/C and IncI1 Plasmids Isolated from Multidrug-Resistant Salmonella Newport Using Single-Molecule Real-Time Sequencing**. *Foodborne Pathogens and Disease* 2018, **15**(6):361-371.
231. Poole TL, Callaway TR, Norman KN, Scott HM, Loneragan GH, Ison SA, Beier RC, Harhay DM, Norby B, Nisbet DJ: **Transferability of antimicrobial resistance from**

multidrug-resistant Escherichia coli isolated from cattle in the USA to E. coli and Salmonella Newport recipients. *Journal of Global Antimicrobial Resistance* 2017, **11**:123-132.
232. Elbediwi M, Pan H, Biswas S, Li Y, Yue M: **Emerging colistin resistance in Salmonella enterica serovar Newport isolates from human infections**. *Emerging Microbes & Infections* 2020, **9**(1):535-538.
233. Zheng J, Luo Y, Reed E, Bell R, Brown EW, Hoffmann M: **Whole-Genome Comparative Analysis of Salmonella enterica Serovar Newport Strains Reveals Lineage-Specific Divergence**. *Genome Biology and Evolution* 2017, **9**(4):1047-1050.
234. Chen C-Y, Strobaugh TP, Jr., Nguyen L-HT, Abley M, Lindsey RL, Jackson CR: **Isolation and characterization of two novel groups of kanamycin-resistance ColE1-like plasmids in Salmonella enterica serotypes from food animals**. *PLOS ONE* 2018, **13**(3):e0193435.
235. Campbell D, Tagg K, Bicknese A, McCullough A, Chen J, Karp BE, Folster JP: **Identification and Characterization of Salmonella enterica Serotype Newport Isolates with Decreased Susceptibility to Ciprofloxacin in the United States**. *Antimicrobial Agents and Chemotherapy*, **62**(7):e00653-00618.
236. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al*: **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"**. *Proc Natl Acad Sci U S A* 2005, **102**(39):13950-13955.
237. de Moraes MH, Soto EB, Salas González I, Desai P, Chu W, Porwollik S, McClelland M, Teplitski M: **Genome-Wide Comparative Functional Analyses Reveal Adaptations of Salmonella sv. Newport to a Plant Colonization Lifestyle**. *Frontiers in Microbiology* 2018, **9**:877.
238. **National Agricultural Statistics Service** [https://quickstats.nass.usda.gov/]
239. Sánchez-Pacheco Santiago J, Kong S, Pulido-Santacruz P, Murphy Robert W, Kubatko L: **Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary**. *Proceedings of the National Academy of Sciences* 2020, **117**(23):12518-12519.
240. Tanner JR, Kingsley RA: **Evolution of Salmonella within Hosts**. *Trends in Microbiology* 2018, **26**(12):986-998.
241. Scott KP: **The role of conjugative transposons in spreading antibiotic resistance between bacteria that inhabit the gastrointestinal tract**. *Cellular and Molecular Life Sciences CMLS* 2002, **59**(12):2071-2082.
242. Aviv G, Rahav G, Gal-Mor O, Davies Julian E: **Horizontal Transfer of the Salmonella enterica Serovar Infantis Resistance and Virulence Plasmid pESI to the Gut Microbiota of Warm-Blooded Hosts**. *mBio*, **7**(5):e01395-01316.
243. Foley Steven L, Kaldhone Pravin R, Ricke Steven C, Han J: **Incompatibility Group I1 (IncI1) Plasmids: Their Genetics, Biology, and Public Health Relevance**. *Microbiology and Molecular Biology Reviews*, **85**(2):e00031-00020.
244. Redondo-Salvo S, Bartomeus-Peñalver R, Vielva L, Tagg KA, Webb HE, Fernández-López R, de la Cruz F: **COPLA, a taxonomic classifier of plasmids**. *BMC Bioinformatics* 2021, **22**(1):390.

245. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, Garcillán-Barcia MP, de la Cruz F: **Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids**. *Nat Commun* 2020, **11**(1):3602.
246. Zaleski P, Wolinowska R, Strzezek K, Lakomy A, Plucienniczak A: **The complete sequence and segregational stability analysis of a new cryptic plasmid pIGWZ12 from a clinical strain of Escherichia coli**. *Plasmid* 2006, **56**(3):228-232.
247. Hooper DC, Jacoby GA: **Mechanisms of drug resistance: quinolone resistance**. *Ann N Y Acad Sci* 2015, **1354**(1):12-31.
248. Oladeinde A, Cook K, Orlek A, Zock G, Herrington K, Cox N, Plumblee Lawrence J, Hall C: **Hotspot mutations and ColE1 plasmids contribute to the fitness of Salmonella Heidelberg in poultry litter**. *PLOS ONE* 2018, **13**(8):e0202286.
249. Folkesson A, Advani A, Sukupolvi S, Pfeifer JD, Normark S, Löfdahl S: **Multiple insertions of fimbrial operons correlate with the evolution of Salmonella serovars responsible for human disease**. *Molecular Microbiology* 1999, **33**(3):612-622.
250. **Almond production in California** [https://apps1.cdfa.ca.gov/FertilizerResearch/docs/Almond_Production_CA.pdf]
251. **California Almond Facts** [https://www.almonds.com/sites/default/files/content/attachments/almond_industry_-_kern_county.pdf]
252. Bedhomme S, Perez Pantoja D, Bravo IG: **Plasmid and clonal interference during post horizontal gene transfer evolution**. *Mol Ecol* 2017, **26**(7):1832-1847.
253. Hughes JM, Lohman BK, Deckert GE, Nichols EP, Settles M, Abdo Z, Top EM: **The role of clonal interference in the evolutionary dynamics of plasmid-host adaptation**. *mBio* 2012, **3**(4):e00077-e00012.
254. Sheppard SK, Guttman DS, Fitzgerald JR: **Population genomics of bacterial host adaptation**. *Nature Reviews Genetics* 2018, **19**(9):549-565.
255. Suzuki M, Doi Y, Arakawa Y: **ORF-based binarized structure network analysis of plasmids (OSNAp), a novel approach to core gene-independent plasmid phylogeny**. *Plasmid* 2020, **108**:102477.
256. Jolley KA, Bray JE, Maiden MCJ: **Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications**. *Wellcome Open Res* 2018, **3**:124.
257. [https://www.cdc.gov/salmonella/newport-07-20/index.html]
258. Bazinet AL, Zwickl DJ, Cummings MP: **A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0**. *Syst Biol* 2014, **63**(5):812-818.
259. **Figtree** [https://github.com/rambaut/figtree]
260. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al*: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing**. *J Comput Biol* 2012, **19**(5):455-477.
261. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics* 2014, **30**(14):2068-2069.
262. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P: **Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper**. *Molecular Biology and Evolution* 2017, **34**(8):2115-2122.

263. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J: **Roary: rapid large-scale prokaryote pan genome analysis**. *Bioinformatics* 2015, **31**(22):3691-3693.
264. Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, Goesmann A: **Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores**. *Microb Genom* 2020, **6**(10):mgen000398.
265. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic acids research* 2004, **32**(5):1792-1797.
266. Tamura K, Stecher G, Kumar S: **MEGA11: Molecular Evolutionary Genetics Analysis Version 11**. *Molecular Biology and Evolution* 2021, **38**(7):3022-3027.
267. Katoh K, Misawa K, Kuma Ki, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic Acids Research* 2002, **30**(14):3059-3066.
268. Steinegger M, Söding J: **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets**. *Nature Biotechnology* 2017, **35**(11):1026-1028.
269. **Snippy** [https://github.com/tseemann/snippy]
270. Chiu JKH, Ong RT-H: **Clustering biological sequences with dynamic sequence similarity threshold**. *BMC Bioinformatics* 2022, **23**(1):108.
271. Levy Karin E, Mirdita M, Söding J: **MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics**. *Microbiome* 2020, **8**(1):48.
272. Meyer F, Hofmann P, Belmann P, Garrido-Oter R, Fritz A, Sczyrba A, McHardy AC: **AMBER: Assessment of Metagenome BinnERs**. *GigaScience* 2018, **7**(6):giy069.
273. Meziti A, Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT: **The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample**. *Appl Environ Microbiol* 2021, **87**(6):e02593-02520.
274. Allen AC, Beck CA, Sattelberger DC, Kiszka JJ: **Evidence of a dietary shift by the Florida manatee (Trichechus manatus latirostris) in the Indian River Lagoon inferred from stomach content analyses**. *Estuarine, Coastal and Shelf Science* 2022, **268**:107788.
275. Lewin Harris A, Robinson Gene E, Kress WJ, Baker William J, Coddington J, Crandall Keith A, Durbin R, Edwards Scott V, Forest F, Gilbert MTP *et al*: **Earth BioGenome Project: Sequencing life for the future of life**. *Proceedings of the National Academy of Sciences* 2018, **115**(17):4325-4333.
276. Delmont TO, Gaia M, Hinsinger DD, Frémont P, Vanni C, Fernandez-Guerra A, Eren AM, Kourlaiev A, d'Agata L, Clayssen Q *et al*: **Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean**. *Cell Genomics* 2022, **2**(5):100123.
277. Tamminen M, Virta M, Fani R, Fondi M: **Large-Scale Analysis of Plasmid Relationships through Gene-Sharing Networks**. *Molecular Biology and Evolution* 2012, **29**(4):1225-1240.
278. Pellow D, Zorea A, Probst M, Furman O, Segal A, Mizrahi I, Shamir R: **SCAPP: an algorithm for improved plasmid assembly in metagenomes**. *Microbiome* 2021, **9**(1):144.

279. Beitel CW, Froenicke L, Lang JM, Korf IF, Michelmore RW, Eisen JA, Darling AE: **Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products**. *PeerJ* 2014, **2**:e415-e415.

280. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang X-S, Davis-Richardson A, Canepa R, Triplett EW, Faith JJ, Sebra R *et al*: **Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation**. *Nature biotechnology* 2018, **36**(1):61-69.