ABSTRACT

Title of Document: WHOLE-GENOME SEQUENCE ANALYSIS FOR PATHOGEN DETECTION AND DIAGNOSTICS

Adam M. Phillippy, Doctor of Philosophy, 2010

Directed By: Professor Steven L. Salzberg
Department of Computer Science

This dissertation focuses on computational methods for improving the accuracy of commonly used nucleic acid tests for pathogen detection and diagnostics. Three specific biomolecular techniques are addressed: polymerase chain reaction, microarray comparative genomic hybridization, and whole-genome sequencing. These methods are potentially the future of diagnostics, but each requires sophisticated computational design or analysis to operate effectively. This dissertation presents novel computational methods that unlock the potential of these diagnostics by efficiently analyzing whole-genome DNA sequences. Improvements in the accuracy and resolution of each of these diagnostic tests promises more effective diagnosis of illness and rapid detection of pathogens in the environment.

For designing real-time detection assays, an efficient data structure and search algorithm are presented to identify the most distinguishing sequences of a pathogen that are absent from all other sequenced genomes. Results are presented that show these "signature" sequences can be used to detect pathogens in complex samples and differentiate them from their non-pathogenic, phylogenetic near neighbors. For microarray, novel pan-genomic design and analysis methods are presented for the characterization of unknown microbial isolates. To demonstrate the effectiveness of these methods, pan-genomic arrays are applied to the study of multiple strains of the foodborne pathogen, *Listeria monocytogenes*, revealing new insights into the diversity and evolution of the species. Finally, multiple methods are presented for the validation of whole-genome sequence assemblies, which are capable of identifying assembly errors in even finished genomes. These validated assemblies provide the ultimate nucleic acid diagnostic, revealing the entire sequence of a genome.

WHOLE-GENOME SEQUENCE ANALYSIS FOR PATHOGEN DETECTION AND DIAGNOSTICS

By

Adam Michael Phillippy

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
2010

Advisory Committee:
Professor Steven L. Salzberg, Chair
Professor Arthur L. Delcher
Professor Mihai Pop
Professor Carleton L. Kingsford
Professor David M. Mosser

Preface

This dissertation is based on the following publications, listed by chapter:

Chapter 2

Phillippy AM, Mason JA, Ayanbule K, Sommer DD, Taviani E, Huq A, Colwell R, Knight T, Salzberg SL (2007) Comprehensive DNA signature discovery and validation. PLoS Comput Biol 3: e98.

Cai M, Phillippy AM, McIver K, Huq A, Salzberg SL, Colwell R, Knight IT. Detection of *Burkholderia pseudomallei* and *B. mallei* using a large-scale, comparative genomics approach. *In preparation*.

Chapter 3

Phillippy AM, Ayanbule K, Edwards NJ, Salzberg SL (2009) Insignia: a DNA signature search Web server for diagnostic assay development. Nucleic Acids Res 37: W229-234.

Chapter 4

Phillippy AM, Deng X, Zhang W, Salzberg SL (2009) Efficient oligonucleotide probe selection for pan-genomic tiling arrays. BMC Bioinformatics 10: 293.

Chapter 5

Deng X[*], Phillippy AM[*], Li Z, Salzberg SL, Zhang W. Probing the pan-genome of *Listeria monocytogenes*: implications for intraspecific niche expansion and genomic diversification. *In preparation*. [*]*Equal contribution*.

Chapter 6

Phillippy AM, Schatz MC, Pop M (2008) Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 9: R55.

Dedication

To my girls, *Katherine* and *Ella*.

I am forever strengthened by your smiles.

Acknowledgements

I owe everything to the two amazing people who first sparked my interest in bioinformatics ten years ago and have guided my scientific career to this point, Art Delcher and Steven Salzberg. My early interactions with them at The Institute for Genomic Research awakened my scientific spirit, and since, they have been outstandingly supportive advisors who have always encouraged me to pursue my interests. I cannot thank them enough for all the wisdom and direction they have given me through the years. I am honored to be their colleague and friend.

Thank you to all my past TIGR colleagues and friends. The wonderful scientific atmosphere at TIGR is what inspired me to pursue my PhD. I appreciate everyone who influenced me and contributed to that decision. Especially, my trusted friend and advisor, Mihai Pop; my dynamic influence, Jacques Ravel; my peaceful influence, Pawel Gajer; my cheerful influence, Rekha Seshadri; my power forward, Dave Rasko; and my wonderful summer fellowship advisor, Jane Carlton. I have benefited greatly, both personally and scientifically, from each of them.

I am lucky to have shared an office with some wonderful people while at the University of Maryland that make both great friends and sounding boards: Mike Schatz, Cole Trapnell, Dave Kelley, Saket Navlakha, James White, Smeds Smedinghoff, and Ben Langmead. Discussions with them made each day at the office a joy. I will miss these daily interactions most of all. A special thanks to my office brother, Mike Schatz, who has shared ideas, lunch, and an office with me for nearly a decade.

I am extremely thankful for my family, especially my wife and daughter whose love I carry with me every day. I am so blessed to have them as my own, and I would not be here without them. To my dad, I owe my unending curiosity and love of discovery—I will forever feel like a kid when I look at the stars. Many thanks to my mom for always keeping me focused and being the world's best Nana. And thanks to my brother and sister for being the most loving siblings I could wish for.

I am deeply indebted to all of my collaborators who contributed to the work presented in this dissertation. I thrive on close and friendly collaborations, and I thoroughly enjoyed working with everyone. Many helped coauthor the publications on which this dissertation is based, and their specific contributions are listed at the end of each chapter. Most notably, Kunmi Ayanbule designed and wrote the beautiful interface for Insignia. Ivor Knight, Jacquline Mason, Lingxia Jiang, and Main Cai designed and carried out the Insignia validation experiments. Rita Colwell, Anwar Huq, Kevin McIver, Elisa Taviani, Henk den Bakler, and Martin Wiedmann contributed isolates and/or DNA samples. Lynn Schriml, Aaron Gussman, and Cesar Arze contributed genome sequence and annotation data for Insignia. Wei Zhang and Xiangyu Deng were very close collaborators on all aspects of the *Listeria* project and contributed valuable biological insight. Michael Schatz and Mihai Pop co-developed the assembly forensics methods, and influenced nearly all parts of this dissertation. Fritz McCall and the entire UMIACS staff provided excellent computational infrastructure support. Finally, many thanks to my entire committee, Steven Salzberg, Art Delcher, Mihai Pop, Carl Kingsford, and David Mosser, for their support and careful reading of this dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Pathogenic microbes, both natural and weaponized, pose significant dangers to human health and safety, yet they are virtually invisible. To detect their presence, assess their threat, and direct treatment, these microbes and their genetic makeup must be revealed through diagnosis. Historical diagnostics have relied upon observation of symptoms and cellular culture. A familiar example is streptococcal infections, which are typically diagnosed from patient symptoms and then confirmed by culture. However, syndromic diagnosis is often inaccurate, leading to improper, missed, or over treatment; and cellular culture is only possible for a very small number of microorganisms. In addition, these methods cannot provide a detailed description of the pathogen, such as its genetic constitution, which is necessary for accurate threat assessment, forensics, or directed treatment. It is predicted that the introduction of improved point-of-care diagnostics could save millions of lives on an annual basis [1] by increasing the efficacy of directed treatment and preventing the over-treatment that breeds resistant microorganisms.

Newly developed genomic approaches offer an alternative diagnostic strategy capable of both rapidly detecting and characterizing all microbes in any environmental or clinical medium with high accuracy. Despite their diversity and adaptability, pathogenic microbes share a common reliance on a nucleic acid genome to survive and function, making it is possible to design tests that will rapidly detect and characterize pathogens based solely on their genomic DNA or RNA. Such tests have a wide range of applications, from diagnosing infections to detecting harmful microbes in the environment. Though relatively new, nucleic acid diagnostics have advanced significantly in the past decade, making them the preferred choice for many applications [2,3]. Future improvements in portability and cost will likely bring these diagnostics into clinical settings. However, due to the large scale of genomic data, and the uncertainties inherent in biochemical reactions, sophisticated computational methods are required to design and analyze these tests.

In addition to clinical applications, pathogen detection and diagnostics have immediate applications to biodefense. For example, in response to the anthrax attacks of 2001, The Institute for Genomic Research was tasked with sequencing the *Bacillus anthracis* attack strain in order to identify its genomic characteristics and point of origin [4]. Shortly after, genotyping microarrays were developed for the rapid and accurate genotyping of *B. anthracis* isolates [5]. Additionally, the U.S. government deployed airborne pathogen monitoring stations at the Salt Lake City Olympic Games and other major metropolitan areas to serve as an early warning for biological attacks [6,7]. These state-of-the-art capabilities were only possible with the aid of advanced bioinformatics.

This dissertation explores computational approaches designed to complement three specific nucleic acid technologies: polymerase chain reaction (PCR), microarray comparative genomic hybridization (CGH), and whole-genome sequencing. Polymerase chain reaction exploits the ability of DNA polymerase to amplify a template DNA sequence into millions of copies. When the template sequence is unique to a particular organism, a successful PCR reaction can signal the presence of that organism's DNA. Comparative genomic hybridization relies on the complementary pairing of DNA to simultaneously test for thousands of short oligonucleotide sequences. CGH requires more time and DNA than PCR, but is capable of characterizing a genome to the level of single nucleotide polymorphisms. Finally, whole-genome sequencing is capable of reading millions of short sequences directly from a genome, and, when combined with genome assembly algorithms, is capable of reconstructing the entire genome of an organism. While the most complete and accurate technique, whole-genome sequencing is currently the most expensive as well. However, rapidly declining sequencing costs will soon make sequencing an important and widely used diagnostic tool.

These three diagnostic methods complement each other and are each applicable to different scenarios. Routine detection and surveillance can be carried out by rapid and portable PCR-based diagnostics, and the more costly and informative diagnostics can be applied only in the event of a detection. For example, a single PCR assay can detect the presence of a given species, but it cannot always determine the specific pathogenic or antibiotic factors present in a genome. Alternatively, for greater cost, microarrays can be designed to probe every known gene of a species to better determine the genomic composition of a detected pathogen. However, all probe-based assays can only test for genomic material that has been included in the assay design. Genome sequencing is the most comprehensive diagnostic tool, capable of decoding the entire genome of a pathogen without any prior assumptions and potentially uncovering novel genes and function.

1.1 Specific Contributions and Organization of the Dissertation

The thesis of this dissertation is that the development and integration of new computational methods with state-of-the-art biotechnology can produce new and more accurate diagnostics for both the detection and characterization of pathogenic microbes. High-performance computing and efficient algorithms are utilized to align, compare, and analyze vast amounts of genomic data in order to design more effective diagnostics. Three types of molecular tests are considered—PCR, CGH, and whole-genome sequencing—each providing a different tradeoff between speed, cost, and detail, and each with particular computational challenges. Chapters 2 and 3 cover methods developed for the design of real-time PCR detection assays. Chapters 4 and 5 cover methods developed for the design and analysis of pan-genomic microarray comparative genomic hybridization experiments. And Chapter 6 covers whole-genome sequence assembly and validation. The specific contributions of this dissertation, organized by chapter, follow:

Chapter 2. I present an efficient method for discovering and storing DNA signatures, which are sequences that distinguish a pathogen's genome from all others. The cornerstone of this method is a novel data structure, called a *match cover*, which captures the essential information necessary to compute DNA signatures. I demonstrate how this data can be computed for all currently available genomes using high-performance computing, and how this structure can be used for rapid signature retrieval from a database using efficient interval set algorithms. I present laboratory validation results for *Vibrio* and *Burkholderia* species signatures, demonstrating the effectiveness of Insignia signatures, outlining a new assay design strategy, and providing a robust set of validated TaqMan PCR assays for the detection of select pathogens.

Chapter 3. I present a Web application that integrates signature discovery, genome annotation, and PCR assay design. This is the first application to make comprehensive signature detection and assay design accessible to investigators lacking high-performance computing resources.

Chapter 4. I introduce the concept of a *Pan-genome Tiling Array* for surveying the gene content of an unknown bacterial isolate using microarray comparative genomic hybridization, and present the *PanArray* algorithm for efficiently designing such arrays. These unique pan-genome tiling arrays are capable of fully tiling all genomes of a species on a single microarray chip, which provides maximum flexibility for CGH analysis.

Chapter 5. I present a novel analysis method for pan-genomic microarrays with superior accuracy and flexibility. I describe a comparative analysis of the species *Listeria monocytogenes* that integrates both genome sequences and microarrays to provide a complete picture of the phylogeny and genetic diversity of the species. To estimate the core genome of the species using both draft sequences and microarrays, I introduce a new mathematical model for core genome regression that better fits the data than prior models. This comparative analysis reveals new insight into the evolution and biology of this important foodborne pathogen.

Chapter 6. I present a taxonomy of common assembly errors and present a suite of methods for validating whole-genome sequence assemblies. I demonstrate how the integrated use of these methods can be used to automatically assess assembly quality. Application of these methods to real assembly data reveals mis-assemblies even in “finished” genomes. These tools will ultimately lead to automated finishing protocols that could dramatically improve the quality of whole-genome assemblies.

Chapter 2

Comprehensive DNA Signature Discovery and Validation[†]

DNA signatures are nucleotide sequences that can be used to detect the presence of an organism and to distinguish that organism from all other species. This chapter describes Insignia, a new, comprehensive system for the rapid identification of signatures in the genomes of bacteria and viruses. With the availability of hundreds of complete bacterial and viral genome sequences, it is now possible to use computational methods to identify signature sequences in all of these species, and to use these signatures as the basis for diagnostic assays to detect and genotype microbes in both environmental and clinical samples. The success of such assays critically depends on the methods used to identify signatures that properly differentiate between the target genomes and the sample background. Insignia computes accurate signatures for most bacterial genomes and makes them available through a Web site. A sample of these signatures has been successfully tested on a set of 46 *Vibrio cholerae* strains and 12 *Burkholderia* spp. strains, and the results indicate that the signatures are highly sensitive for detection as well as specific for

[†] This chapter includes previously published work with multiple authors. See Section 2.7 for details.

discrimination between the target strains and their near relatives. The Insignia method, whereby the entire genomic complement of organisms are compared to identify probe targets, is a promising method for diagnostic assay development, and provides assay designers with the flexibility to choose probes from the most relevant genes or genomic regions. The Insignia system is freely accessible via a Web interface at: http://insignia.cbcb.umd.edu.

2.1 Background

Modern health and security concerns have raised interest in the real-time detection and identification of pathogenic microbes. Bacterial and viral pathogens have always represented one of the greatest threats to human health, and in recent times this threat increased due to the possibility of engineered biological agents. For these and other reasons, the genome sequencing field has targeted and sequenced the complete genomes of hundreds of bacteria and thousands of viruses over the past decade, with many more sequences expected to appear in the near future. These sequences now make it possible to develop probe-based assays capable of identifying any of hundreds of organisms in environmental and clinical samples. Such assays rely on detecting a DNA sequence that distinguishes the target organism from all other known bacteria and viruses, and from background material, which could include DNA from humans, other animals, plants, or other species. A probe that accurately distinguishes between a target genome—or set of genomes—and all other background genomes is termed a signature sequence.

By this definition, a signature sequence must be conserved among a set of target genomes and dissimilar to any sequence in the surrounding environment. To detect a target with existing technology such as qPCR assays, signatures must be relatively short; however, if they are too short, they will not be unique. For example, because there are only $4^{10} \approx$ 1 million 10 bp (base-pair) sequences, and a typical bacterial genome is more than 1 million bp in length, most 10-mers will be shared by many genomes and therefore make unsuitable signatures. Increasing the length, k, of the signature alleviates this problem, but if k is too large, it may not be possible to find a signature shared by a set of target genomes. Therefore, there is a tradeoff between signature sensitivity (the number of genomes that share the signature) and specificity (the number of genomes that do not possess the signature). For instance, a long signature may be highly specific to a particular strain or isolate, but it may not be sensitive enough to detect closely related strains that might cause the same disease or have other shared phenotypic characteristics. Because genomic sequence is nonrandom, and only a small sample of genomes has been sequenced, it is difficult to estimate an optimal signature length. In practice, signature length is usually determined by the constraints of the detection technology (e.g., ~20 bp for PCR primers).

Current probe-based technologies are generally based on either PCR or microarray hybridization. These methods are beginning to replace traditional gel-based fingerprinting because they can more effectively differentiate between closely related microbes [8]. Microarray methods are particularly promising because of their ability to multiplex many probes on a single chip [8,9,10], improving both the redundancy and capabilities of the diagnostic. PCR does not multiplex as nicely; however, it remains popular because of its robustness, speed, and low cost [11,12,13].

Unlike restriction fingerprinting, both PCR and microarray methods require explicit knowledge of the underlying DNA sequence, therefore necessitating probe design.

Traditional probe design strategies have focused on single genes or other loci that are determined *a priori* to be useful in distinguishing one target organism from another. Examples include genes that are associated with phylogenetic distance (e.g., 16S rRNA genes) and variable number tandem repeats (VNTRs). In the former case, where the gene or locus is conserved among target and non-target organisms, gene sequence alignments would be used to aid in probe design. Probes would then be manually designed and screened for sensitivity and specificity to the target. Those assays failing to identify all target organisms, or producing false positives, would be invalidated and the design revised. This manual screening made diagnostic assay design expensive and only worth doing for a few select pathogens. Alternatively, variable number tandem repeats (VNTRs) have proven very useful in classifying and distinguishing many closely related strains of bacteria, such as *Bacillus anthracis* whose 16S rRNA sequences are identical [14,15]. Although these methods are effective, they only provide a limited number of signatures, which are not always sufficient to identify bacteria or viruses in a new sample; in particular, if the sample contains an unknown strain, it might contain genetic variability in precisely the region for which assays are designed. Thus, in general, one would like to have as many assays available as possible. Insignia addresses this by using the complete genome to generate all unique signatures, from which the assay designer can choose those that are best suited for a particular application.

Recent increases in the amount of available genomic sequence have made it possible to largely automate the design and screening of probes via computational search algorithms. Large-scale computational prediction of DNA signatures was first undertaken for the Biological Aerosol Sentry and Information System (BASIS), deployed at the Salt Lake City Olympic Games in 2002 [6,7]. The related BioWatch project operates by collecting and analyzing airborne microbial samples for known pathogens, using PCR probe-based detection methods. Newer aerosol detection systems, such as the Autonomous Pathogen Detection System (APDS) [16], automate the process, and can identify a known bioweapon in 0.5 to 1.5 hours [17]. Similar techniques are not limited to aerosols, and can be used in clinical or agricultural settings [18].

The success of these assays depends on both the available sequence databases and the computational methods used to identify signatures that differentiate the threat organisms from the background. Signature design for both the Biological Aerosol Sentry and Information System (BASIS) and BioWatch was handled by Lawrence Livermore National Laboratories (LLNL), and what began as a simple proof-of-concept BLAST search at LLNL evolved into the sophisticated KPATH signature pipeline [19]. KPATH identifies sequences shared by a collection of target genomes, yet unique with respect to all other microbial genomes, and is notable for its ability to handle such a large search space. Other methods for probe selection more rigorously address hybridization efficiency (binding energy, self-hybridization, etc.), but do not scale well for large target and background sets [20,21,22,23]. Most notable are the approaches that promise the scalability of KPATH combined with the hybridization considerations of the other methods [24,25].

Because of its history of use in real-world diagnostic systems, a more detailed description of KPATH is warranted. It consists of four major components. First, a whole-genome multi-alignment is performed on a set of target genomes. This produces a “consensus gestalt,” which represents the sequences that are conserved in all the target genomes. Next, this consensus is matched against a database of background sequences using Vmatch [26]. This step computes all exact matches between the target consensus and the background. Matching sequences are masked out to create a “uniqueness gestalt,” which represents all sequences that are shared between target genomes and unique with respect to the background. Third, signature sequences are supplied to the Primer3 program [27], which designs PCR assays based on those sequences. Primer3 produces a set of oligos suitable for testing by a TaqMan PCR assay: a forward primer, a reverse primer, and an intervening probe oligomer [28]. Finally, assay candidates are screened using BLAST [29] for near matches that might disrupt the hybridization process, and ranked according to their satisfaction of PCR experimental constraints. The result of this four-stage process is a set of ranked, prescreened assays, which are then subjected to rigorous laboratory validation. The transition to these computational methods from previously manual design methods has resulted in greatly increased design efficiency by limiting the number of assays that fail during laboratory validation.

While highly innovative, the KPATH pipeline is not publicly available, and many of the sequences and signatures remain secret. In addition, KPATH requires significant computing resources (hours of computing time on a 24-CPU server [19]), which are beyond the means of many investigators. In contrast, Insignia is a transparent, highly accessible signature pipeline, with the entire system being controlled by a Web interface and all supporting software released under an open source model. Additionally, Insignia dramatically accelerates the discovery process by pre-computing exact sequence matches for all genomes and storing this information in a specialized data structure for rapid retrieval.

Using the Insignia Web interface, users select a desired signature length and a set of target genomes. After query submission, the system analyzes the stored match information, and identifies signature candidates in less than one minute. Candidates may then be further screened using experimental constraints (melting temperature, GC content, etc.), or using further computational criteria, such as the existence of near matches that may cause cross-hybridization. The integrated Gemina database (http://gemina.tigr.org), which includes detailed annotation and supplementary epidemiological information for major pathogens, provides further support for signature selection. This rich metadata allows the formulation of complex queries such as “find signatures shared by all enteric *Escherichia coli*,” and it allows the user to search for signatures in the context of the surrounding annotation. Insignia can compute signatures for any microbial genome in GenBank (both draft and complete), and screens signatures against a comprehensive background including all bacterial, archaeal, and viral sequences, plus additional eukaryotic sequences from National Center for Biotechnology Information (NCBI) RefSeq database [30]. The Insignia

Web interface is fully described in Chapter 3. The following sections describe the principal computational components and validation of the results.

2.2 *Signature Discovery Pipeline*

Given a set of target genomes, a set of background genomes, and a signature length *k*, Insignia identifies all *k*-mer signatures present in the target genomes. A *k*-mer signature is a string of *k* nucleotides that is perfectly conserved in a set of target genomes, but does not occur exactly in any of the background genomes. A short *k*-mer (e.g. 15–20 bp) is more likely to be shared by a group of target species, but is also more likely to appear in the background. When Insignia finds a series of *k*-mer signatures each overlapping by *k*-1 bases, it reports these longer chains as a single region, where every *k*-mer in the chain is guaranteed to be absent from the background. Occasional background *k*-mer matches will occur by chance, but long chains of *k*-mer signatures are likely to represent the sequences most dissimilar from the background. This can be thought of as the inverse of typical seed-and-extend alignment strategies such as BLAST [31]. Converse to assuming similarity exists in regions sharing exact matches, users of Insignia can assume that dissimilarity exists in regions devoid of exact matches.

Signature chains are reported as an interval from the start of the first signature word to the end of the last signature word in the chain. The signature chain $[s,e]$ contains exactly $(e - s - k + 2)$ signature words of length *k*, completely covering the interval $[s,e]$ in the target sequence. Signature words are, by definition, perfectly conserved in all target genomes, and contain at least a single difference from every background sequence. (Note that a signature may occur multiple times in a target genome; it is not required to occur just once). Therefore, a signature chain will contain a difference from any background genome at least every *k* bases. For some types of detection assays, a difference every ~20 bp is not sufficient for discrimination. However, polymorphisms tend to be unevenly distributed, and similar sequences are likely to share at least one exact match over a long distance. In the validation studies, long signature chains (e.g. >100 bp) follow this tendency and are often quite dissimilar from the background. After identifying candidate signatures, a more sensitive BLAST search of the background can be performed to identify any similar but non-identical matching sequences.

Signature chains have the benefit of being both long and specific. Thus, they make ideal targets for PCR-based detection assays such as TaqMan, which work well with an amplicon length of around 100 bp and specific primers and probes. For microarrays, signature chains can be tiled across their length with multiple probes to provide adequate redundancy. These two techniques can be combined to achieve a very high degree of accuracy. PCR primers can be designed at the boundaries of a long signature chain, and the interior of the chain can be tiled with microarray probes. The detection procedure could then consist of PCR amplification, followed by microarray hybridization or sequencing of the product. Alternative probe-based detection strategies, such as melting curve analysis [32], are also available that are sensitive to a single nucleotide difference in the probe sequence.

Insignia provides real-time signature retrieval for an arbitrary set of target and background genomes. This requires the vast majority of computational work be done in advance and cached, so that a minimum amount of computation is necessary at the time of the query. To accommodate this, Insignia is designed as two separate components: the match pipeline and the signature pipeline. This distinction separates the computationally intensive matching step from the much simpler signature generation step, and allows sequence matches to be recomputed offline as new genomes become available. While the matches may take days to compute, the signatures can be extracted from this cached information in seconds. From the website, users select a set of target genomes and a set of background genomes, and an exhaustive set of signatures is computed and displayed in seconds.

2.2.1 Match Pipeline

The function of the match pipeline is to identify exact matches between all pairs of target and background sequences in the database. The size of the Insignia sequence database is currently more than 100 billion nucleotides, and even with the linear-time algorithms described below, this is too large to search in real time. Limiting targets to microbial genomes saves some computational effort, but the process of matching all pairs of target and background genomes remains expensive. A schematic of the match pipeline is shown in Figure 2.1.

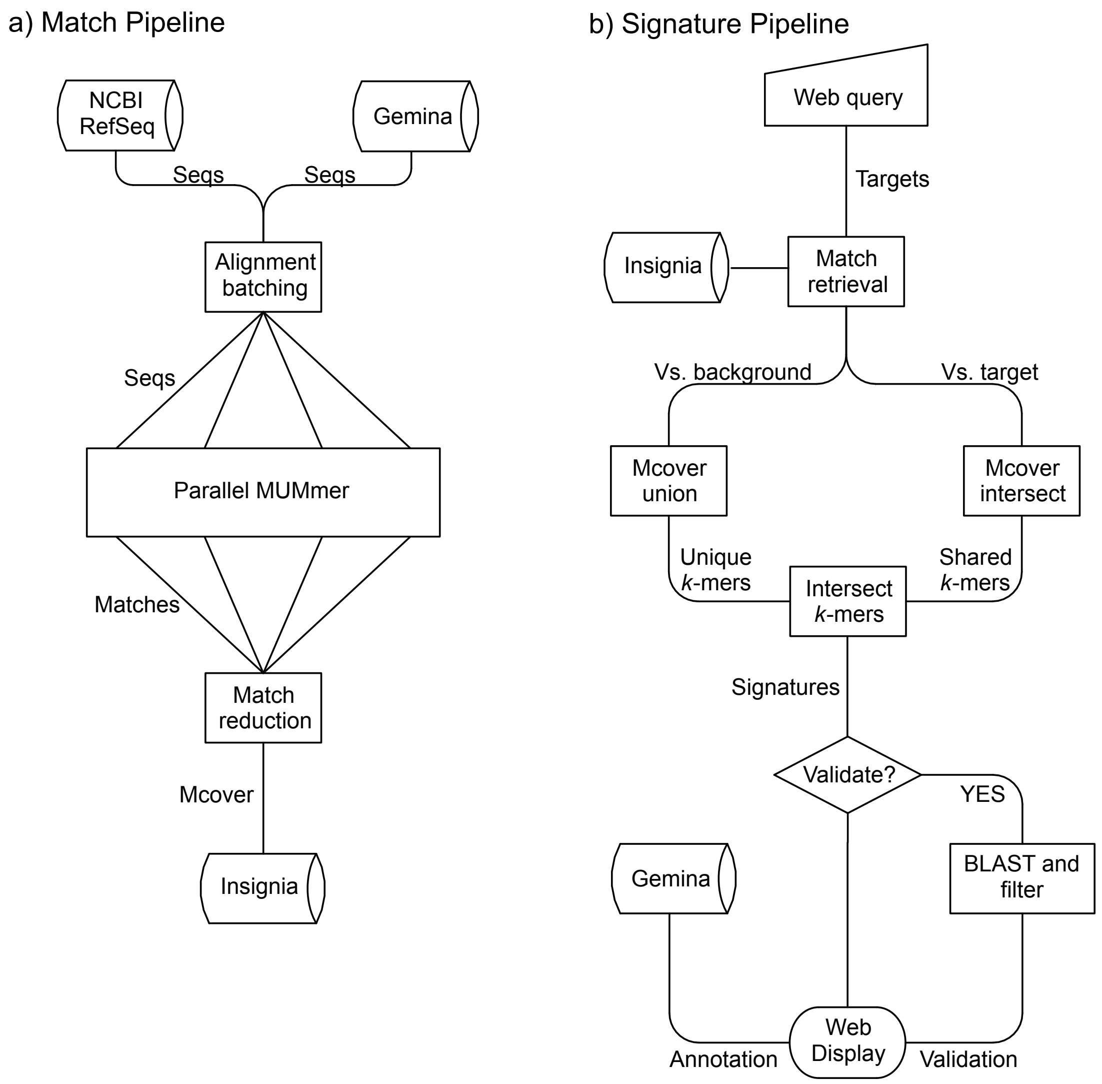


Figure 2.1: Schematic of the Insignia signature pipelines a) Exact matches are computed using MUMmer on a distributed cluster. Matches are converted to the reduced match cover format and stored in a database for quick retrieval. b) Users select a set of target and background genomes on the website, and match information for these genomes is retrieved from the Insignia database. Unique and shared *k*-mers are computed from the match cover data in parallel, and these results are intersected to identify signatures that are shared by the targets and absent from the background.

2.2.1.1 Maximal Exact Matches

To complete the matching phase within a reasonable amount of time, all maximal exact matches of 18 bp or longer are first identified between all pairs of database sequences. A maximal exact match (MEM) is a perfect match between two substrings than cannot be extended in either direction without encountering a mismatch. These are efficiently computed using MUMmer [29–31], a linear time and space suffix tree matching algorithm. To expedite the process, MUMmer searches are partitioned across a 192-node Linux cluster. Even with the use of an efficient search algorithm, however, the size of the database and the high repeat content of many genomes cause the size of the output—the number of matches between all pairs of genomes—to reach unmanageable levels (e.g., the number of matches can be quadratic with respect to the size of the genomes). To combat this problem, matches are converted to a minimalized *match cover* data structure, described next. This structure saves space by scaling linearly and later provides a convenient mechanism for computing signatures.

2.2.1.2 Match Cover Data Structure

The *match cover*, M_{tb}, of a target genome *t*, with respect to some background genome *b*, is simply the list of intervals on *t* that are covered by contiguous, exact matches to genome *b*. To eliminate redundancy, all intervals contained within larger intervals are removed, but overlapping intervals are not merged. This assures that every subinterval matches contiguously to some portion of the background sequence, and every maximal match to the background is contained by a single interval (Figure 2.2). After construction of the match cover, the intervals are sorted by their start position, and stored as a list of (*start*, *length*) pairs. Because this structure only stores the target "half" of the match data, space requirements are reduced by eliminating irrelevant background match coordinates. What remains is a minimal set of intervals on genome *t* that exactly match some part of genome *b*.

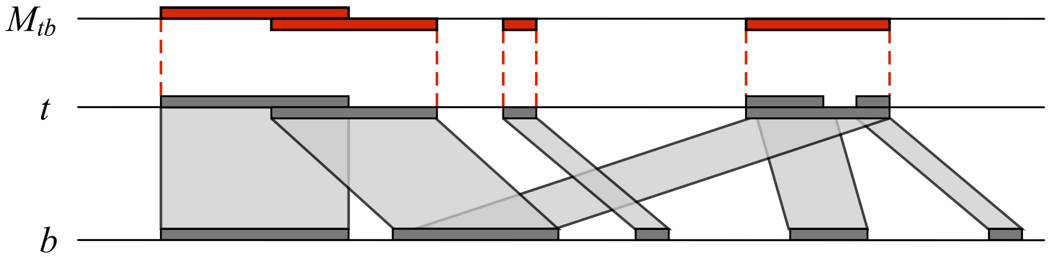


Figure 2.2: The match cover data structure (M_{tb}) is shown for the exact matches between a target (t) and background (b) genome. M_{tb} intervals (red boxes) represent regions of the target with a contiguous match to the background (gray boxes).

By storing only the location of the matches on the target genome, the match cover also eliminates redundant information caused by repetitive sequences. Take for instance, two potential target genomes t and u. Because all target genomes are, by default, part of the background, two match covers will be created, M_{tu} and M_{ut}. Now assume an identical repeat occurs x times in t, and x times in u. A list of exact matches (*start t*, *start u*, *length*) would require $3x^2$ integer values to represent the repeat, while the combined match covers would require only $4x$ values. Therefore, even when storing both halves of a match set ($t \rightarrow u$ and $u \rightarrow t$), the match cover scales linearly, rather than quadratically, making it more efficient in dealing with repeats. This behavior was empirically tested for an all-versus-all comparison of ~300 bacterial genomes, and the match cover reduced the match list from its original size of 78 GB to just 2 GB. This 39-fold space reduction demonstrates the prevalence of repetitive matches in real data and the utility of the match cover structure. Considering the match cover is simply a list of intervals, standard data compression could be applied to obtain further space savings.

The match cover is not a lossless conversion, however, because it discards information about where a match occurred in the background. The remaining information is nonetheless sufficient for signature computation, where it suffices to know only which regions of a target are unique. Furthermore, because it scales linearly, large background databases can be accommodated without drastically increasing the match cover size, and draft quality genomic sequences can be incorporated without difficulty. As the next section will show, the match cover encapsulates all the necessary information for signature discovery and allows for the rapid construction of signatures for any set of target and background genomes in linear time.

For perspective, it is worth mentioning that the match cover is an equivalent, interval representation of matching statistics [33,34]. Both formalizations represent the longest contiguous match beginning at any position of a sequence, but an interval representation is more space-efficient and easier to interpret in the context of signature discovery. Rahmann also leverages the properties of matching statistics in describing a "jump list" for the discovery of DNA probes [25], and it is interesting to note that although the match cover and jump list were arrived at independently, they are analogous given their shared utilization of matching statistics.

2.2.2 Signature Pipeline

The function of the signature pipeline is to generate valid signatures for any set of target and background genomes. Given a set of target genomes, a set of background genomes, and a signature length parameter *k*, Insignia identifies all *k*-mer signatures present in the target genomes. Because there are thousands of possible targets and many more backgrounds, combinatorics rules out the pre-computation of all signatures; however, it is possible to generate signatures from the match information with minimal overhead. The pipeline for doing so is divided into two parallel stages, corresponding to the two primary criteria a valid signature must meet: 1) a signature must be shared by all genomes in the target set; and 2) a signature must not exist in any genome in the background set.

2.2.2.1 Shared Sequences

To determine which *k*-mers are shared between a set of target genomes, one target is chosen as the reference *r*, and all match covers, M_{rt}, are intersected for each *t* in the target set using a plane sweep algorithm on the sorted interval lists. This intersection yields all matches shared by the target genomes relative to the sequence of the reference genome. Given the resulting match cover intersection I_r for a collection of targets, a *k*-mer in *r* is shared by all other target genomes if, and only if, it is entirely contained within a single interval of I_r (Figure 2.3). To avoid returning signatures containing ambiguous base codes, shared sequences are only permitted to contain the characters {A,C,G,T}.

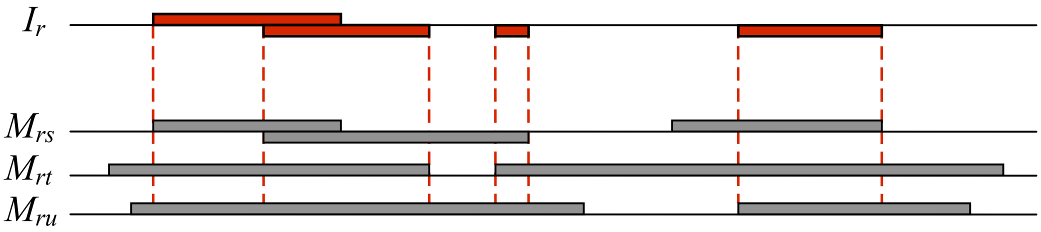


Figure 2.3: Computing shared sequences obtained from an intersection (I_r) of three match covers M_{rs}, M_{rt}, M_{ru}. I_r intervals (red boxes) represent regions of the reference *r* shared with all other target genomes *s*, *t*, *u* as derived from the match covers between the reference and each target (gray boxes).

2.2.2.2 Unique Sequences

A parallel stage of the signature pipeline computes a list of *k*-mers unique to a target genome with respect to some background. Once again, the match cover information is leveraged to efficiently identify these *k*-mers. Assuming the same target reference *r*, all match covers, M_{rb}, are merged for each *b* in the background set. Because each interval list is sorted by position in *r*, this can be done efficiently by simply merging the sorted lists. This produces a consolidated set of matches to the reference from the background. Maximal matches smaller than *k*, and matches entirely contained by another interval, are irrelevant and can be removed. Given the resulting match cover union U_r for a collection of backgrounds, a *k*-mer in *r* is unique with respect to the background if, and only if, it is not entirely contained within a single interval of U_r (Figure 2.4). It is sufficient to compute unique *k*-mers with respect to a single target, because a sequence will only be reported as a signature if it is also shared by all target genomes. Thus, any single target is guaranteed to contain all of the ensuing signature sequences.

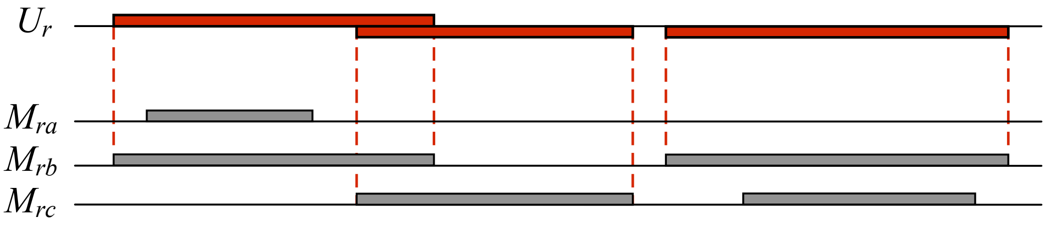


Figure 2.4: Computing unique sequences obtained from a union (U_r) of three match covers M_{ra}, M_{rb}, M_{rc}. U_r intervals (red boxes) represent regions of the reference *r* matching some background genome *a*, *b*, *c* as derived from the match covers between the reference and each background (gray boxes).

2.2.2.3 Signature Computation

The interval set operations for processing match covers are extremely efficient. For M_{RT} sorted reference-target match intervals and T target genomes, the time complexity for finding shared-mers is $O(M_{RT} \log T)$, with the log component incurred by the plane-sweep's priority queue of overlapping interval endpoints. In practice, the number of target genomes is few and the complexity is dominated by the size of the match list. The time complexity for computing unique-mers by merging the sorted interval sets is linear, $O(M_{RB})$ for M_{RB} reference-background intervals. The results of these two operations are then intersected again to identify sequence signatures, i.e. *k*-mers that are both shared by the targets and unique with respect to the background. Therefore, the complexity of extracting signatures from a match cover database is linear with regard to the number of reference match intervals. For a typical target and background set, this translates to under one minute of processing, given the current database size and processing speeds.

2.2.3 Implementation

The computational demanding components of Insignia are implemented in C++, with driver scripts written in Perl. The Insignia database uses indexed flat files. The Insignia code is freely released open-source at: http://insignia.cbcb.umd.edu.

2.3 *Vibrio cholerae TaqMan Validation Pilot Study*

As a pilot validation study, Insignia was used to develop assays for the identification of *Vibrio cholerae* at the species level using a TaqMan Real-Time qPCR format. The initial version of Insignia queried a database that was populated with ~300 bacterial genomes, including one strain of *V. cholerae* (O1 biovar El Tor strain N16961), and four near neighbors in the family Vibrionaceae (three *Vibrio* and one *Photobacterium* species). Thus the question for Insignia was: among all available DNA sequences, what sequences are unique to *V. cholerae*? The Insignia Web interface was used to retrieve all 20-mers unique to *V. cholerae*, from which 50 TaqMan assays were designed from 50 randomly selected signature chains of varying lengths.

To test whether the signature assays were broadly inclusive of *V. cholerae* strains, the 50 assays were tested against a panel of 46 strains of *V. cholerae* comprising a global distribution of both clinical and environmental strains from all major serotypes. To test whether they excluded non-cholera vibrios, the assays were additionally tested against a panel of 22 nearest-neighbor species in the family Vibrionaceae, along with one *Escherichia coli* control. Figure 2.5 and Figure 2.6 show example inclusive and exclusive qPCR results, respectively.

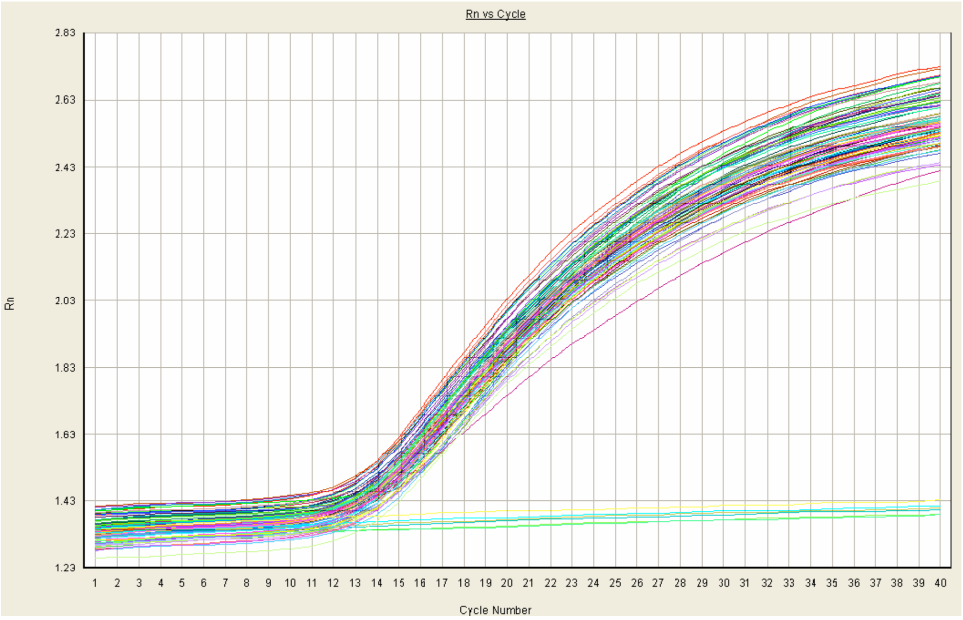


Figure 2.5: Example inclusive TaqMan assay displaying increased fluorescence due to target amplification for all 46 *V. cholerae* strains tested, and no fluorescent activity among the *E. coli* negative controls. Relative florescence intensity (y-axis) for 40 PCR cycles (x-axis) is shown.

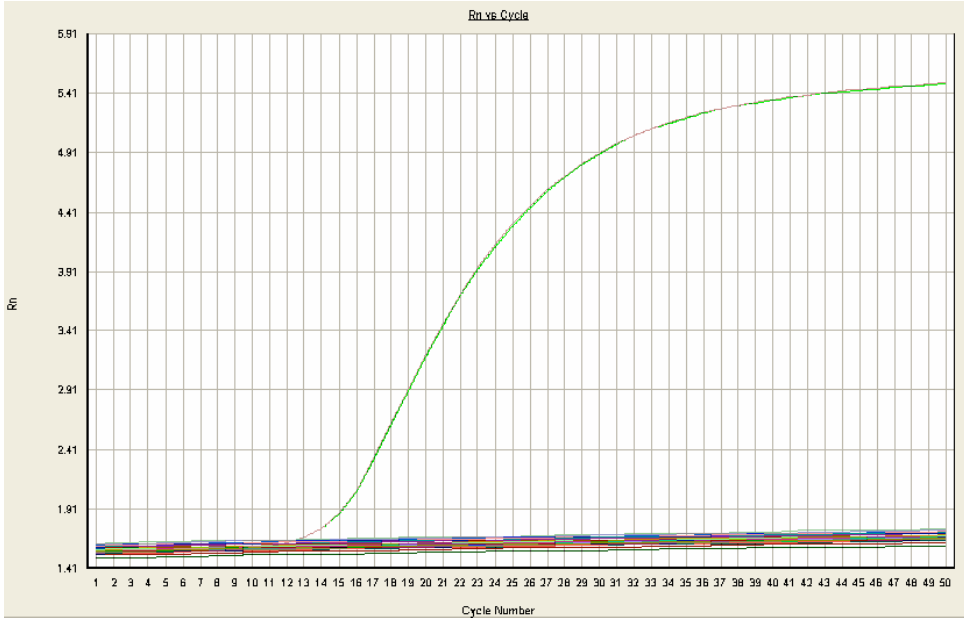


Figure 2.6: Example exclusive TaqMan assay displaying increased fluorescent activity for the reference strain of *V. cholerae* and no fluorescent activity among the 23 non-cholera strains. Relative florescence intensity (y-axis) for 40 PCR cycles (x-axis) is shown.

2.3.1 *Vibrio* Assay Design and Validation

The nucleotide sequences of the probes and primers for each TaqMan assay were selected from the signature set identified by Insignia for *Vibrio cholerae O1 biovar El Tor strain N16961*. The probes and primers were designed outside of Insignia using commercially available design software (Allele ID, Premier Biosoft International, http://www.premierbiosoft.com). All assays were designed for PCR to run under the same conditions.

All PCR assays were conducted in duplicate and Ct values were used to evaluate the extent to which each assay was inclusive of *V. cholerae* strains and/or excluded near neighbor strains. Ct values of ≤20 were considered strong positive, and Ct values between 20 and 50 were binned in increments of 4 (i.e., 21–24, 25–28, etc.) to simplify analysis of the relative efficiency of PCR across all assays and strains.

2.3.2 *Vibrio* Validation Results

Figure 2.7 summarizes the validation results for the 50 assays, covering 69 organisms, and totaling 3,450 experiments. Each square in Figure 2.7 represents one experiment, with color indicating the qPCR Ct value (the number of PCR cycles before amplification is detected). Green and yellow squares indicate relatively rapid amplification while orange and red indicate delayed or failed amplification. As the figure makes clear, most assays detected all *V. cholerae* strains, with approximately half of the assays providing strong detection capability for every one of these diverse strains. The effectiveness of some assays deteriorated slightly for the non-O1/O139 serotypes, although they still provided positive results. This was to be expected, however, given that only a single *V. cholerae* strain (of serotype O1) was available to Insignia at the time of design. Additional genomic sequences from the other serotypes would have undoubtedly removed many of these less-efficient signatures from the Insignia output. Gardner *et al.* explore this phenomenon further in the context of viral signature development [35].

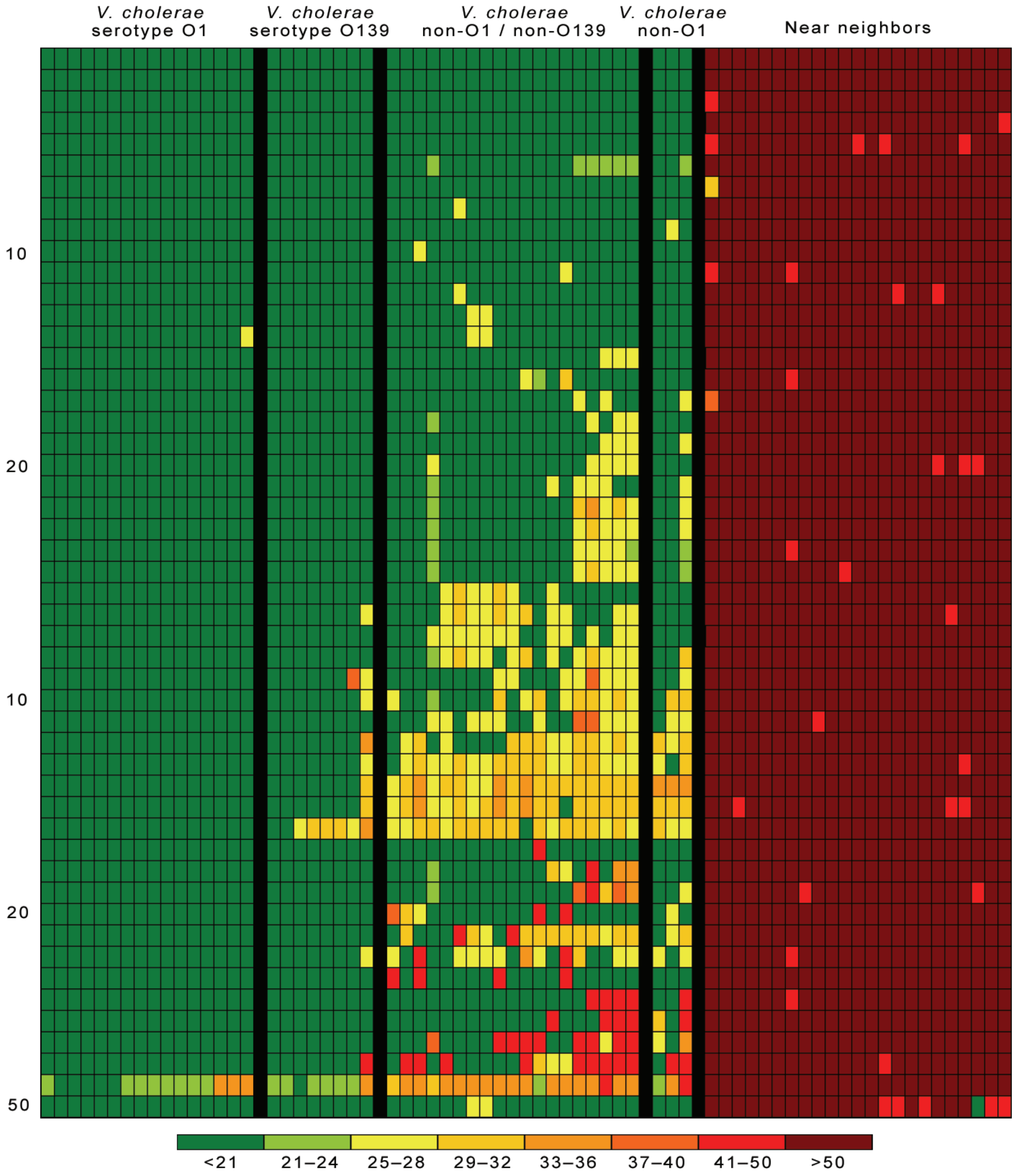


Figure 2.7: *Vibrio cholerae* assay validation results for the 50 assay designs tested on 46 *V. cholerae*, 22 near neighbors, and one *E. coli* control. Organisms (columns) are grouped by serotype, and assays (rows) are sorted vertically by effectiveness. Each colored box represents the Ct value for one of the 3,450 validation experiments. For example, assays 1–5 show strong amplification for all *V. cholerae* strains and heavily delayed or failed amplification for all other organisms.

In addition to successful detection of a wide variety of *V. cholerae* strains, all but one of the tested assays (98%) were able to successfully discriminate between *V. cholerae* and its near neighbors. Furthermore, 1,115 of the 1,150 exclusive tests (97%) had Ct values >50, indicating that all of the tested *V. cholerae* signatures are either absent or significantly divergent from the other members of Vibrionaceae. Assay signature sequences, inclusive and exclusive strain information, and detailed qPCR results for all validation experiments are available from the Insignia website.

2.4 <u>Burkholderia pseudomallei and B. mallei Discrimination Study</u>

For a more thorough analysis of signature performance, Insignia was used to develop 100 TaqMan PCR assays aimed at detecting *Burkholderia pseudomallei* and/or *Burkholderia mallei* in clinical and environmental samples. Unlike the traditional approach, which targets one or two specific *Burkholderia* genes, Insignia's large-scale comparative genomics approach objectively identifies the most suitable assay designs. 10 of these assays were designed and validated for detecting *B. pseudomallei*, 10 for detecting *B. mallei*, and 80 for detecting either one.

The validation results show that 88% of the computationally designed assays yield little or no false-positive signal for near-neighbor strains and have 100% sensitivity to a panel of target strains, with a detection limit of 1.5 to 15 equivalent genomes per PCR assay. Furthermore, all 20 species-specific assays were 100% sensitive and 100% specific among the strains tested. A duplex PCR assay for identifying and discriminating *B. pseudomallei* and *B. mallei* simultaneously was also developed and tested. The benefits of utilizing large-scale comparative genomics for designing PCR detection assays are highlighted, and 88 new, validated, and robust *Burkholderia* detection assays are now available.

2.4.1 Introduction to *Burkholderia*

Burkholderia pseudomallei and *Burkholderia mallei* are the etiologic agents of melioidosis and glander [36,37,38,39], respectively, two severe infectious diseases in animals and humans. The two pathogens are listed as Category B select agents by the Center for Disease Control and Prevention, due to their ease of dissemination, high morbidity, and high mortality rates [40]. Although *B. pseudomallei* and *B. mallei* differ in physiology, ecology, and epidemiology, they are phylogenetically close [41,42]. Both species have two chromosomes [43,44], and their 16S rDNA sequences are almost identical [41]. It has been proposed that *B. mallei* be classified as a subspecies of *B. pseudomallei*.

Rapid and accurate identification of these two pathogens—including discrimination from non-pathogenic *Burkholderia* species—is critical for early diagnosis and effective management of a potential bioterrorism attack. Previously, diagnostic tests for rapid detection, identification, and discrimination of these two pathogens and their non-pathogen neighbors have used a traditional approach that identifies, *a priori*, candidate genes for developing differentiating assays. Candidates include 16S rDNA [41,45,46], 23S rDNA [47], flagellin (*fliP* and *fliC*) [45,46,48], metalloprotease gene (*mprA*) [49], ribosomal protein subunit S21 (*rpsU*) [46], type III secretion systems (TTS1 and TTS2) [50,51], and locus P27 [52], locus 8653 and 9438 [53]. These approaches requires either a prior understanding of the phylogenetic relationship between the target pathogens and non-target organism or an understanding of the molecular mechanisms of pathogenesis of the target pathogens.

Approaches targeting phylogenetically relevant loci are complicated by the fact that *B. pseudomallei* is relatively heterogeneous, with single nucleotide polymorphisms being common, even in highly conserved loci, such as 16S and 23S rDNA and other housekeeping genes [42]. In addition, *B. pseudomallei* and *B. mallei* share high sequence similarity with *Burkholderia thailandensis* [54] and other *Burkholderia* species. Therefore, it is difficult to find consensus sequences in phylogenetically relevant genes that are fully representative, yet not found in near neighbors. As a result, most assays reported in the literature require two or more individual assays to identify and discriminate *B. pseudomallei* and *B. mallei*, and their close neighbors.

This section presents the Insignia strategy for identifying signature sequences, which examines entire genomes for suitable signatures and does not require knowledge of the phylogeny or molecular pathogenesis of the target pathogens. Using this approach, Insignia uncovered genomic signatures differentiating the two pathogens, *B. pseudomallei* and *B. mallei*, from all other known sequence, as well as signatures differentiating the two pathogens themselves.

2.4.2 *Burkholderia* Assay Design and Validation

The Insignia Web interface was used to generate all *Burkholderia* signatures. At the time of this study the list of available targets included 6 *B. pseudomallei* strains, 3 *B. mallei* strains, and 1 *B. thailandensis* strain. To identify *B. pseudomallei* specific signatures, the 6 *B. pseudomallei* strains were selected as the target and compared with all other genome sequences as background. To identify *B. mallei* specific signatures, the 3 *B. mallei* strains were selected as the target. To identify signatures common to both *B. pseudomallei* and *B. mallei*, all 9 combined *B. pseudomallei* and *B. mallei* strains were selected as the target. The signature *k*-mer length was set to 20 bp for all queries, and the background was set to include all available sequenced genomes (both draft and finished). At the time of this study, the Insignia background contained 3,801 bacterial, archeal, and viral genomes, and all of the NCBI RefSeq genomic DNA, containing the genomes of human, animals, and plants (RefSeq release #23). In addition, the default Insignia result filters were used to remove signature chains whose average melting temperature (Tm) was lower than 45° C and average GC% was lower than 30%.

Only those signature chains extending to 99 bp or longer were chosen for development of TaqMan assays. 100 TaqMan assays were developed from 100 signature chains using Primer3 [27], of which 10 were *B. pseudomallei* specific, 10 were *B. mallei* specific, and 80 were common to both *B. pseudomallei* and *B. mallei*. The PCR product size range was specified at 75 bp to 150 bp. The optimum melting temperature (Tm) of primer was set at 60 °C, and the optimum Tm of probe was set at 70 °C. All primer and probe sequences were subjected to BLAST searches against the NCBI *nr* database to confirm their uniqueness to the target strains.

A duplex TaqMan assay to distinguish *B. pseudomallei* from *B. mallei* was developed by combining a *B. pseudomallei* specific assay with a *B. mallei* specific assay, but with separate dyes. The 4 corresponding primers, 2 probes, and template

DNA were added into the TaqMan assay at the same final concentrations as in the single TaqMan PCR assays.

2.4.3 *Burkholderia* Validation Results

Insignia identified a total of 45,279 signatures chains unique to *B. pseudomallei* and 11,748 unique to *B. mallei*. A further 197,538 chains were identified as shared by both *B. pseudomallei* and *B. mallei*, but unique to these two species compared to background. Chains long enough to accommodate PCR targeting were less frequent, but were, nonetheless, frequent enough to provide at least 100 candidates for assay design (Table 2.1). The physical positions of the 100 signatures that were used for TaqMan assay development on the chromosomes of *B. pseudomallei K96243* and *B. mallei ATCC23344* are shown in Figure 2.8.

Table 2.1: Summary of identified *Burkholderia* signatures.

	Number of signatures found for		
Signature chain length	***B. pseudomallei* and *B. mallei***	***B. pseudomallei***	***B. mallei***
20–98	196,297	44,626	11,723
99–149	1,052	525	17
150–198	131	98	3
≥199	58	30	5
Totals	**197,538**	**45,279**	**11,748**

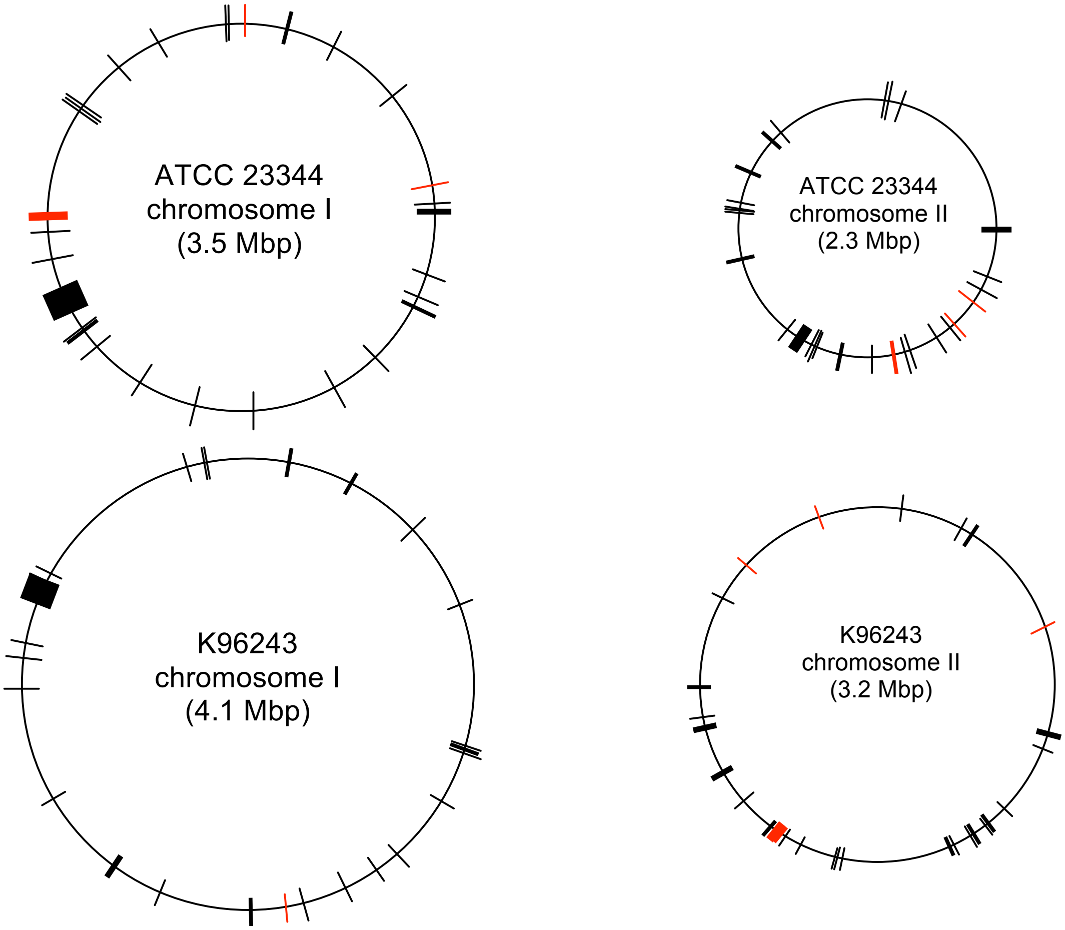


Figure 2.8: Location of *Burkholderia* signature chains used for assay development on chromosome I and/or II of *B. mallei* (ATCC 23344) and *B. pseudomallei* (K96243). Positions in black are signatures common to both, while positions in red are species-specific signatures.

The results of the 100 TaqMan assays are summarized in Figure 2.9. The 10 *B. pseudomallei*-specific assays (assays 81–90) showed low cycle threshold Ct values (15–19) for all 6 *B. pseudomallei* strains tested and no detectable signal with *B. mallei* or any other *Burkholderia* strains tested. The 10 *B. mallei*-specific assays (Assays 91–100) showed low Ct values (15–19) for all 3 *B. mallei* strains and no detectable signal with *B. pseudomallei* or other near neighbors tested, except for one strain, *B. pseudomallei K96243* (BEI NR-9320). NR-9320 showed a detectable signal at very high Ct values (between 30–34) for all 10 *B. mallei* specific assays. For the 80 assays targeting for both *B. pseudomallei* and *B. mallei*, 68 assays identified all the target strains with very little or no cross reactions with their near neighbor strains, and the Ct values ranged from 15–21. Of the failed assays, ten failed to detect 1 or 2 target strains, and 2 assays cross-reacted with elevated Ct values with their closest neighbor strain *B. thailandensis ATCC 700388*. In summary, 88 out of 100 assays (88%) correctly differentiated all targets from their near neighbors, and for the 20 species-specific assays both the sensitivity and specificity of detection was 100%. Assay signature sequences, inclusive and exclusive strain information, and detailed qPCR results for all validation experiments are available from the Insignia Web site.

The detection limits for all 20 *B. pseudomallei* and *B. mallei* specific assays were in the range of 0.01 pg to 0.1 pg genomic DNA per reaction, equivalent to approximately 1.5 to 15 genomes per reaction, respectively. The presence of 15 ng competitor DNA in the TaqMan assay did not change the detection limits.

The duplex PCR identified the target strain correctly at Ct values of 16–20 with 15 ng template DNA per assay. No obvious interferences between the two assays were detected in the tests except for BEI NR-9320 (K96243), which cross-reacted with *B. mallei* assay at a Ct value of 38.0. The same Ct values were obtained while *B. pseudomallei* and *B. mallei* DNA were co-present at 15 ng for each in the assay.

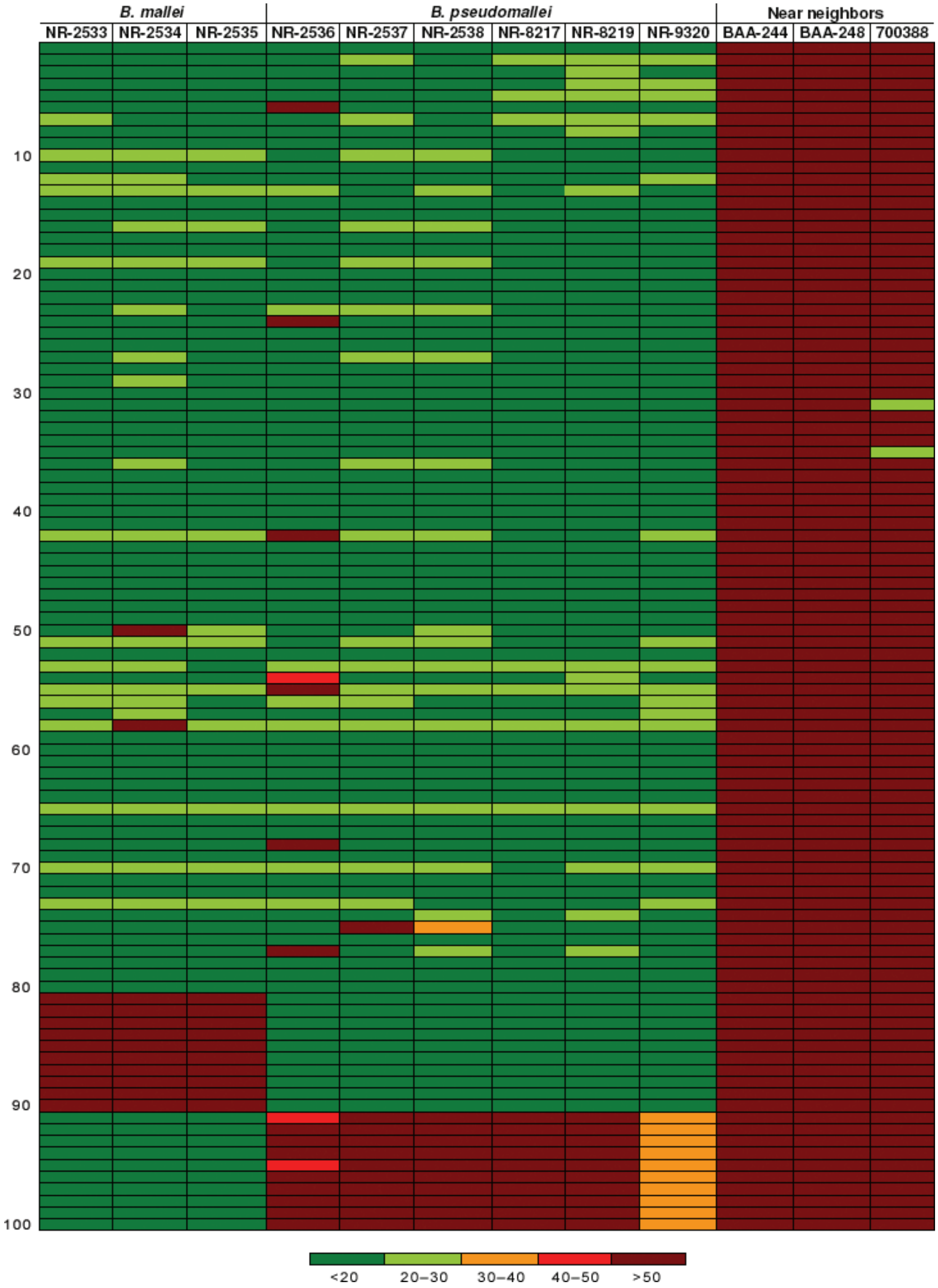


Figure 2.9: *Burkholderia* assay validation results for 100 assays (rows) run using *B. mallei, B. pseudomallei,* and near neighbor strain DNA samples (columns). Again, Ct values are color-coded to indicate the effectiveness of each assay, low Ct values in green and high in red.

2.4.4 Discussion of *Burkholderia* Detection Assays

An ideal diagnostic assay should be able to recover the entire population of target isolates that might be encountered in either clinical or environmental samples, while discriminating between true targets and near neighbors. Insignia's whole-genome, comparative strategy is able to rapidly identify targets for genetic assay design that meet these criteria. Validation using pathogenic *Burkholderia*, as an example, shows this approach is also easily translated into new assays of defined specificity.

In the validation experiments, 88 of the 100 signature assays achieved 100% sensitivity for the strains tested. Of the 12 assays that failed to achieve 100% sensitivity, 2 assays (assay 50 and 58) failed to identify *B. mallei* NR-2534, 7 assays (assays 6, 24, 42, 54, 55, 68, and 77) failed to identify *B. pseudomallei* NR-2536, and one assay (assay 75) failed to identify *B. pseudomallei* NR-2537 and NR-2538. All four of the strains that were not detected by these assays (NR-2534, NR-2536, NR-2537, and NR-2538) have not yet been sequenced and, therefore, were absent from the Insignia database. Without genomic information for all strains, failed assays are to be expected, especially for strain NR-2536, perhaps the most phylogenetically distant of the 6 *B. pseudomallei* strains included in this study. Without genomic sequence data for all strains, validation employing a large panel of test strains is particularly important.

2.4.4.1 Cross Reactivity

Slight cross-reaction was observed for near neighbors of *B. pseudomallei* NR-9320 (K96243) and *B. thailandensis* 700388 (E264). While still easily distinguishable from the target strains, NR-9320 showed a Ct value of 30–34 in all 10 *B. mallei* specific assays and 700388 a Ct value of 26–30 in 2 assays (assays 31 and 35). To validate the original Insignia designs, the primer and probe sequences for these assays were queried against the genomes of *B. pseudomallei K96243* and *B. thailandensis 700388* using Vmatch [26], allowing an edit distance of 4 for each oligo (mismatches and indels). For the cross-reacting NR-9320 assays, no potential products were detected where the primers and probe aligned within 20 Kbp of each other, in correct order and orientation, in the finished K96243 genome sequence. This search also failed for the raw sequencing reads of the K96243 project obtained from the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov).

Among the six *B. pseudomallei* strains tested, NR-9320 was the only one that showed slight cross-reaction with *B. mallei* specific assays. To test for possible *B. mallei* DNA contamination in the *B. pseudomallei* DNA preparation, fresh NR-9320 DNA, with the same lot number, was prepared and tested, with identical results. Sequencing of the amplification product from two rounds of conventional PCR, using the assay primers and *B. pseudomallei* NR-9320 template DNA, revealed that the 134 bp amplicon was identical to *B. mallei 23344* sequence. This sequence match and the fact that all 10 of the *B. mallei*-specific assays cross reacted with *B. pseudomallei* NR-9320 at similar Ct values, taken together with the fact that both DNAs were prepared in the same facility, suggests low level contamination of the *B. pseudomallei* NR-9320 DNA preparation with *B. mallei* DNA.

For assays that cross-reacted with B. thailandensis 700388 (assays 31 and 35), regions of similarity to *B. mallei* were found in the targeted loci. In all cases, however, Insignia properly identified regions of variability between the species, and each primer and probe contained at least one SNP with regard to 700388. DNA sequence harboring the whole amplicon of 700388 of assay 42 was confirmed by DNA sequencing. Computational prediction and filtering of potentially cross-reacting probes is planned as a future improvement to Insignia.

B. pseudomallei NR-2536 DNA also showed slight cross-reactivity (Ct 36–38) with *B. mallei* specific assay 91 and 95 in one of the duplicated tests, while the other test consistently yielded a negative result (Ct >50). Using the same amount of *B. mallei* DNA, the Ct values were 16–18. The large difference in Ct values (average difference of 20) between assays, using the known positives and those using *B. pseudomallei* NR-2536 indicates at least six orders of magnitude difference in sensitivity. Nonetheless, the cross-reactivity was repeatable, suggesting NR-2536 DNA possesses corresponding sequences of assay 91 and 95, with variations at the site where primers and probe bind, resulting in poor amplification.

2.4.4.2 Repetitive Signatures Improve Sensitivity

Assays designed from repetitive sequences will have lower detections limits. For example, assay 75 was developed from a 192 bp signature chain contained in a gene encoding for ISBma2 transposase. It was later discovered that there are 46 copies of this sequence in *B. mallei ATCC 23344* (36 copies in chromosome I, and 10 in chromosome II), and 5 copies in *B. pseudomallei K96243* (4 copies in chromosome I, and 1 in chromosome II). 4 of the 6 *B. pseudomallei* validation strains and 3 of *B. mallei* validation strains possess this sequence. Genomes with this signature registered a lower Ct value of about 11–13 in *B. mallei* and 15–16 in *B. pseudomallei*, compared with the Ct value of 16–20 for the other assays. Given the assay detection limit of about 1.5 to 15 copies of genomic DNA per reaction, the assay targeting this signature should be able to identify some *B. mallei* and *B. pseudomallei* strains from only a single copy of the genome.

2.4.4.3 Effect of Newly Sequenced Draft Genomes

The whole-genome comparative strategy depends on availability of genomic sequence for both target organisms and near neighbors. *Burkholderia* is well suited for this strategy because it is a genus with many sequences available. At the time of assay design, 13 fully sequenced genomes of the *Burkholderia* species were in the public domain and able to be queried by the Insignia pipeline. Shortly after the assay validation was finished, the number of sequenced *Burkholderia* increased to 50. To assess the effect the newly sequenced genomes would have on signature detection, the signature computations were run again, with the results compared with the original designs. With only 6 *B. pseudomallei* and 3 *B. mallei* genomes originally in the database, 1,241 signature chains ≥99 bp in length were identified for co-detection of *B. pseudomallei* and *B. mallei*. Targeting the same 9 genomes and, with the addition of 15 new near-neighbor genomes, this number was reduced to 750 chains ≥99 bp. Currently, with 21 *B. pseudomallei*, 10 *B. mallei*, and 19 near-neighbor genomes in the database, Insignia returned no signature chains ≥99 bp.

After investigating the sharp decline, it was determined that many of the new genomes were low quality draft sequences, which are missing large chunks of their genomes. These gaps make it impossible to find long stretches of perfectly conserved sequence between the targets. Of the 22 new *B. pseudomallei* and *B. mallei* genomes, 10 had assemblies of 1,000 contigs or more. After removing these assemblies from the analysis and concentrating on the 21 higher quality genomes (11 *B. pseudomallei*, 10 *B. mallei*), 438 signature chains ≥ 99 bp are identified, demonstrating the importance of high quality sequences for identifying signatures. The largest reduction in signature candidates was not caused by addition of near-neighbor genomes, but by addition of low quality target genomes. The updated signature set, though smaller than in the original, will be sensitive to a more diverse set of target strains and more specific, because the search was against a much larger panel of near neighbors.

The degree of conservation or divergence among the target strain population also will affect signature detection. The more homologous the target strains, the more reliable the resultant assays. For example, 10 *B. pseudomallei* specific signatures and 10 *B. mallei* specific signatures were evaluated by aligning the sequences with 20 newly sequenced *B. pseudomallei* and 10 *B. mallei* sequences. These data showed the *B. mallei* strains to be relatively homologous, while the *B. pseudomallei* strains were relatively heterogeneous, results consistent with previous studies [41,42]. Accordingly, *B. mallei* specific assays can be expected to be more inclusive than *B. pseudomallei* assays. Fortunately, the designs from this study were based on 6 *B. pseudomallei* genomes (as opposed to 3 for *B. mallei*), which helped the design process compensate for increased heterogeneity by locating the most conserved regions of the genome. In this study, multiple genomes were essential for success, whereas in the prior study of the more homogeneous *Vibrio cholerae*, only a single sequenced target genome was sufficient for a favorable design.

2.5 Future of Insignia

Insignia outputs signature candidates, rather than high confidence, laboratory-validated signatures. However, the validation studies demonstrate that most of these candidates can work quite well as laboratory assays. Due to the limited availability of genomic sequence in public databases (relative to the diversity of all organisms), and the possibility of near-match cross-hybridization, it is difficult to validate a genomic signature via purely computational methods. Instead, Insignia provides a computational screening regimen that eliminates many invalid signatures, so that laboratory validation may focus on the most likely candidates. Additional sequencing will help overcome the computational limitation, and future work on Insignia will be focused on improving the quality of the signatures produced.

In addition to the computational restrictions, limitations of TaqMan PCR have been demonstrated for rapidly diverging target genomes, such as hepatitis and HIV viruses [35,55]. However, for typical bacterial targets, TaqMan assays remain one of the most rapid and sensitive methods for signature detection. In the case where TaqMan is inadequate, different detection technologies, such as chip hybridization, could be used to remove the TaqMan requirement for three adjacent probes and to provide greater signature redundancy. Insignia would easily support the design of such assays.

Viruses pose significant challenges for all detection methods because of their small genomes and high mutation rates. The Insignia database contains thousands of viral genomes; however, for large target sets there are often no conserved signatures. To address highly divergent targets, future Insignia versions may include the ability to identify signatures with degenerate bases, for cases where no exact signature is shared between them. An alternative is to compute the minimum signature set, where each signature might not identify every target, but the set contains at least one identifying signature for each target. This requires clustering the target sequences by similarity and then designing multiple, non-degenerate assays to target each individual cluster. This approach is particularly suited for chip assays where signatures can be multiplexed. A related approach selects combinations of non-unique probes, such that certain viral strains can be identified by their hybridization pattern [56].

Finally, the current version of Insignia can identify signature sequences with a mismatch every 18 bases to the background, but certain types of assays, such as microarrays, require a greater distance between the signature sequence and the background. Signatures containing at least 2 or 3 mismatches to the background may be downloaded from a separate interface for each species group, but are not provided for all combinations of target genomes like the 1-mismatch signatures. Work is planned to expand the current algorithms to identify these more divergent signatures for any combination of target genomes. A possible solution would reduce the search space by identifying all 1-mismatch signatures using the current algorithms, and then apply a more sensitive search algorithm to these candidate signatures to guarantee suitable uniqueness.

2.6 Summary

The validation results indicate that whole genome signature discovery, whereby the entire genomic complement of organisms are compared to identify probe targets, is a promising new tool for diagnostic assay development. A key difference between the comparative genomics approach and a more traditional genetic approach to assay design is that the latter is focused on finding signatures within well defined regions of the genome, while the former utilizes all of the genetic content of the target organism. This provides more flexibility in choosing candidate assays for validation, does not require *a priori* knowledge of the role and function of the candidate loci, and increases the chances of designing a successful assay. Insignia also achieves unmatched scale by screening all microbial genomes against a comprehensive background, while maintaining rapid access to DNA signatures through its Web interface.

To date, hundreds of the discovered signatures have been experimentally validated using TaqMan PCR assays for the detection of multiple pathogens, including *Vibrio cholerae*, *Francisella tularensis*, *Burkholderia mallei*, and *Burkholderia pseudomallei*. The validation studies have revealed that prioritizing the signature chains by length is an effective strategy, and the validated signatures have shown very little cross-reaction with near-neighbor species. Insignia signatures have also been used for microarray genotyping and detection assays. In all cases, the Insignia signatures were shown to be highly sensitive for detection as well as specific for discrimination between near relatives and environmental backgrounds.

2.7 Author Contributions

A version of this chapter appeared previously in published form:

Phillippy AM, Mason JA, Ayanbule K, Sommer DD, Taviani E, Huq A, Colwell R, Knight T, Salzberg SL (2007) Comprehensive DNA signature discovery and validation. PLoS Comput Biol 3: e98.

A version of the *Burkholderia* section is in preparation for publication:

Cai M, Phillippy AM, McIver K, Huq A, Salzberg SL, Colwell R, Knight IT. Detection of *Burkholderia pseudomallei* and *B. mallei* using a large-scale, comparative genomics approach. *In preparation*.

I developed and implemented all computational methods and authored the text. The *Vibrio* validation results were coauthored with Jacquline Mason and Ivor Knight. The *Burkholderia* validation results and discussion were coauthored with Main Cai and Ivor Knight. Jacquline Mason and Main Cai performed the validation experiments. Kunmi Ayanbule engineered the Insignia Web interface and managed the Insignia mySQL database. Rita Colwell, Anwar Huq, Kevin McIver, and Elisa Taviani helped determine the validation strains and contributed the DNA samples. Lynn Schriml, Aaron Gussman, and Cesar Arze contributed genome sequence and annotation data. Steven Salzberg helped conceive the problem, guided the project, and edited the text.

Chapter 3

Web Application for Signature Search and Assay Development[†]

Insignia is made available as a Web application for the rapid identification of unique DNA signatures. These signatures can be used as the basis for diagnostic assays to detect and genotype microbes in both environmental and clinical samples. Insignia identifies an exhaustive set of accurate DNA signatures for any set of target genomes, and screens these signatures against a comprehensive background that includes all sequenced bacteria and viruses, the human genome, and many other animals and plants. Identified signatures may be browsed by genomic location or proximal genes, filtered by composition, viewed in a genome browser, or directly downloaded. Integrated PCR primer design is also provided for each signature. The Insignia website is free and open to all users (http://insignia.cbcb.umd.edu).

3.1 *Background*

Insignia provides a convenient Web interface for identifying genomic signatures from a database of all current bacterial and viral genomic sequences, which

[†] This chapter includes previously published work with multiple authors. See Section 3.4 for details.

currently comprises nearly 14,000 distinct organisms. The input to Insignia is any set of target and background genomes selected from the online database, and the output is a list of signatures perfectly conserved by all target genomes and absent from all of the background genomes. Insignia is the only Web application capable of performing this whole-genome signature design, whereby entire genomes are screened against a comprehensive background. KPATH [19] and TOFI [57] perform similar computations, but must be run offline and require considerable computational resources. To quickly identify signatures for any combination of target and background genomes, and enable use over the Web, Insignia maintains a specialized database containing pre-computed matches between every pair of genomes. Using this match information, signatures are computed "on the fly" in a matter of seconds, using the efficient interval set operations described in the previous chapter. The following sections describe the associated Web application, which provides a convenient interface for signature search and assay design.

3.2 Interface

The Insignia home page provides a full listing of all target genomes currently in the database, organized alphabetically. The database mirrors the Gemina database, maintained by the Institute for Genome Sciences [58]. Currently, the Gemina database includes 13,928 distinct organisms, comprising 11,274 viruses and 2,653 bacteria. This list includes multiple genomes for many well-known species. For instance, only a single *Vibrio cholerae* genome sequence existed in 2006, when the original validation study was performed, but the database now contains 21 *V.*

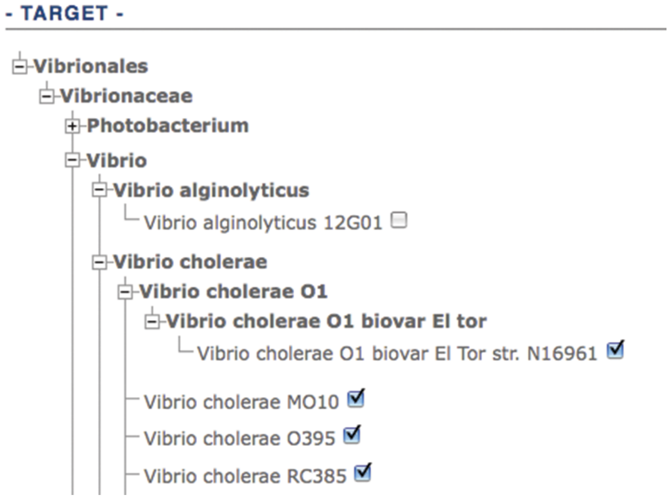
cholerae genomes. The database is updated routinely, and the accuracy of the predicted signatures will continue to increase as more genomes are added.

General help information and examples are also linked from the home page. Links are provided to automatically run signature searches for some important pathogens. Additional links are provided in the "Recent Searches" box after custom searches are performed. This allows users to rerun a previous search with identical parameters. This is helpful for users who may return to the same results many times over the course of designing an assay.

In addition to displaying the current database status and search links, the Insignia home page also hosts a collection of validated signatures. These signatures were predicted by Insignia and have undergone laboratory validation against a series of samples including phylogenetic and environmental neighbors. Validated signatures are currently available for *Vibrio cholerae*, *Francisella tularensis*, *Burkholderia mallei*, and *Burkholderia pseudomallei*. Additional validation studies for *Brucella* spp. and *Yersinia pestis* are planned.

3.2.1 Search Page

The Insignia Search page is available under the "Run Insignia" link on the home page. On the Search page, users are asked to select a reference genome, a set of target genomes, a background, and a signature word length. The reference genome is a member of the target set that provides the genome coordinates for purposes of reporting signatures, and that is used for displaying signatures along with genome annotation. Therefore, it is preferable to select a finished and annotated genome as the reference. The reference genome may be selected from a list of all genomes in the database, or users may type the name of the genome in the search box and the list will be automatically filtered for genomes matching that name. Additional target genomes may be selected from a list in a similar way as the reference genome, or selected from a taxonomy tree of closely related organisms. Figure 3.1 shows a partial view of the target tree for a *Vibrio cholerae* reference. The tree view makes it easy to select all strains from a species or group of related species.



- TARGET -

- ⊟ **Vibrionales**
  - ⊟ **Vibrionaceae**
    - ⊞ **Photobacterium**
    - ⊟ **Vibrio**
      - ⊟ **Vibrio alginolyticus**
        - Vibrio alginolyticus 12G01 □
      - ⊟ **Vibrio cholerae**
        - ⊟ **Vibrio cholerae O1**
          - ⊟ **Vibrio cholerae O1 biovar El tor**
            - Vibrio cholerae O1 biovar El Tor str. N16961 ☑
        - Vibrio cholerae MO10 ☑
        - Vibrio cholerae O395 ☑
        - Vibrio cholerae RC385 ☑

Figure 3.1: Insignia taxonomic selection tree selector showing the order Vibrionales. Subtrees can be expanded or collapsed by clicking, and checkboxes are provided for adding specific genomes to the target set.

After a reference and target have been selected, users may modify the background set if necessary. By default, the background includes all genomes in the database, except for the genomes that have been selected as targets. Buttons are available to exclude draft genomes from the background, or to exclude draft genomes in the same genus as the reference. Draft genomes may contain large gaps, and including them in the target set may have the unintended consequence that no shared signatures can be found; for this reason, it is best to ignore draft genomes unless they are known to be nearly complete. Genomes with >100 contigs may interfere with signature computations and should probably be excluded from the target set and from the background if they are of the same species as the target. The interface provides a simple check box to allow the user to exclude all draft genomes from the background.

Insignia is primarily designed for detecting bacterial signatures, but it can equally well design signatures for viruses. Viruses tend to have much smaller genomes and much higher mutation rates (especially RNA viruses), with the consequence that for a given viral species, there might not be any shared signatures among the sequenced isolates. However, if any signatures exist for a set of viruses, Insignia is guaranteed to find them. To simplify usage for bacterial assay design, all viral genomes can be hidden from view with a checkbox, but they will remain in the background unless specifically excluded.

The final option on Search page is to select a signature word length. Insignia is capable of identifying k-mer signatures for any $k \geq 18$, but as the size of the background continues to increase, many k-mers will hit the background simply by chance and reduce the length of the resulting signature chains. A small k is desirable because it increases the minimum frequency of differences between a signature chain and the background. For the current database size, $k = 20$ appears to be a reasonable compromise between long signature chains and small k. Insignia was tested with $k = 20$ and a background including all of the NCBI RefSeq genomes [30], >100 billion nucleotides, and signature chains >100 bp were still identified for most bacterial species. In future versions of Insignia, if the signature chains are insufficiently long for a PCR product, users will have the ability to design assays across multiple, neighboring signatures.

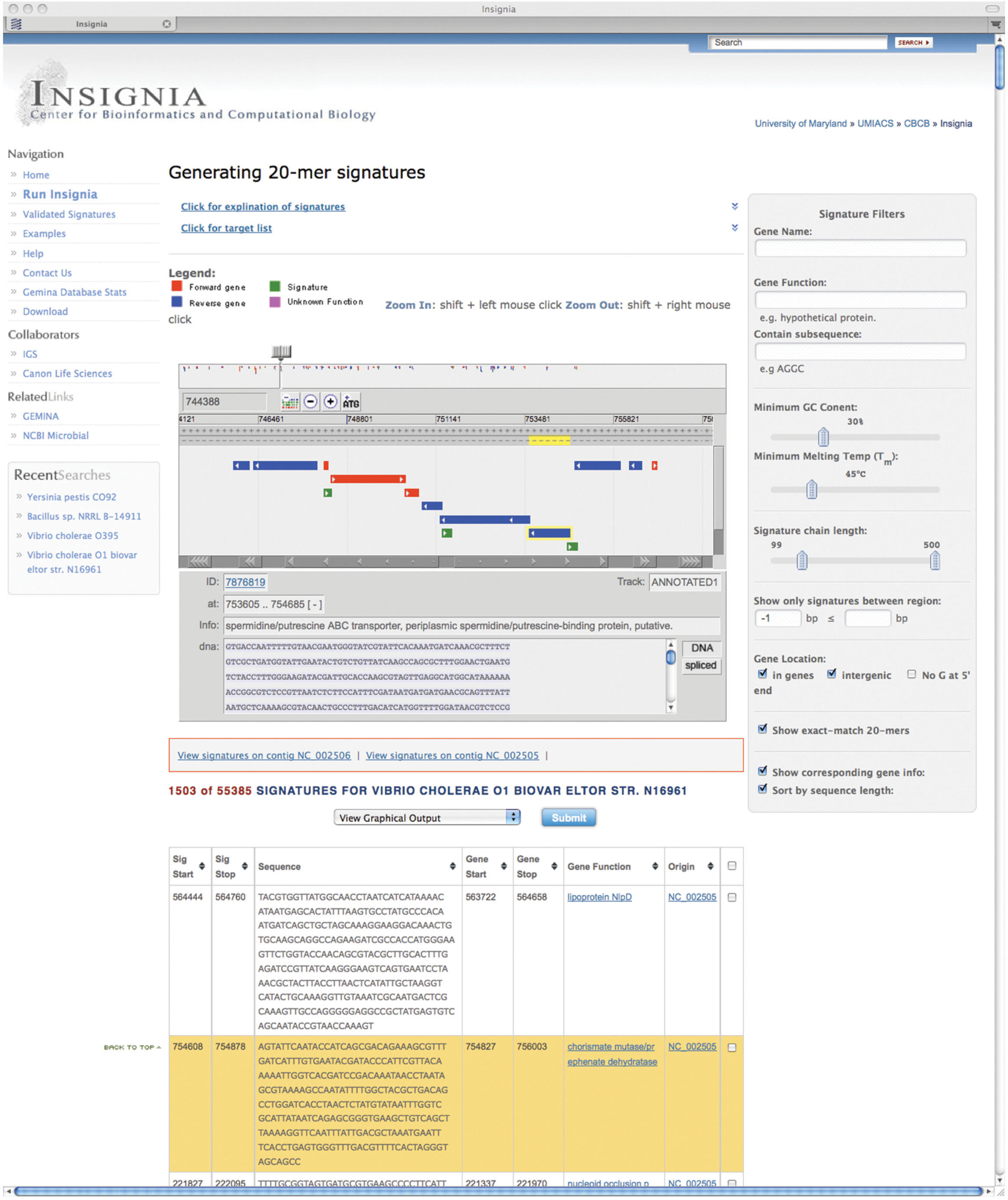
Insignia also provides species-specific, *e*-unique signatures for most reference genomes. An *e*-unique signature is a *k*-mer signature that requires at least (*e* + 1) string edits (substitutions, insertions, deletions) to match the background of other species. These pre-computed signatures are specific to the reference genome, but are not necessarily conserved among other genomes of the same species. If they coincide with an exact-match signature for that species, than they are both *e*-unique and conserved throughout the species. As with the exact-match signatures, adjacent *e*-unique signatures are displayed as chains. If *e*-unique signatures are available for the selected reference genome, a checkbox appears next to the selection box, giving users the option of including these signatures in the search results. 1- and 2-unique species-specific signatures for word lengths 18 and 19 bp are currently available for most finished bacterial genomes via the Web interface, and 1-unique signatures for word lengths 20 through 25 are available upon request.

In addition to the Search page, Insignia accepts query submission via URLs to support links from external sites. The URL format contains the requested signature word length, and the GenBank taxonomy identifiers (TaxID) of the target genomes. The first TaxID is taken as the reference genome, and everything else as the target. If the first TaxID is an internal taxonomy branch, all children are included as the target and the user will be requested to select a reference. For example, to find 20-mer signatures for all strains in the species *Bacillus anthracis*, the URL is (http://insignia.cbcb.umd.edu/results.php?len=20&taxid=1392), where "1392" is the GenBank taxonomy identifier for the species *B. anthracis*. Additional TaxIDs can be added as a comma separated list.

The Gemina website uses this URL interface to link to Insignia (http://gemina.igs.umaryland.edu). Gemina is a database of epidemiology metadata linked to the genomes of infectious pathogens. Through this website, users can search for pathogens via their associated metadata such as transmission method, hosts, symptoms, or geographical location. After a group of organisms with the desired traits are selected on the Gemina website, an Insignia search can be launched to identify genomic signatures for those organisms.

3.2.2 Results Page

Figure 3.2 shows the Results page of the Insignia website after a signature search has completed. The target of this signature search was all 17 strains of *Vibrio cholerae*, with all other genomes as the background. At this stage, all signatures can be downloaded in bulk for further analysis, or browsed on the Results page. The center of the page displays the resulting signature chains (hereafter referred to only as "signatures"). By default, a table of signatures is displayed with start and stop positions, and the signature sequence. The left of the page provides links to relevant websites, help information, and recent searches. The right of the page provides dynamic JavaScript controls for filtering the signatures. In the page displayed in Figure 3.2, 55,385 signatures have been identified but the filter has been set to show only those $\geq$ 100 bp in length, which left 1,503 for display. Using the bottom two checkboxes, annotation information was added to the table and the signatures were sorted in order of decreasing length. In addition, the graphical display has been enabled by choosing "View Graphical Output" from the Submit menu in the center of the page.



Generating 20-mer signatures

Click for explination of signatures

Click for target list

Legend:
🟥 Forward gene 🟩 Signature
🟦 Reverse gene 🟪 Unknown Function

Zoom In: shift + left mouse click **Zoom Out:** shift + right mouse click

ID: 7876819
Track: ANNOTATED1
at: 753605 .. 754685 [-]
Info: spermidine/putrescine ABC transporter, periplasmic spermidine/putrescine-binding protein, putative.
dna:

```
GTGACCAATTTTTGTAACGAATGGGTATCGTATTCACAAATGATCAAACGCTTTCT
GTCGCTGATGGTATTGAATACTGTCTGTTATCAAGCCAGCGCTTTGGAACTGAATG
TCTACCTTTGGGAAGATACGATTGCACCAAGCGTAGTTGAGGCATGGCATAAAAAA
ACCGGCGTCTCCGTTAATCTCTTCCATTTCGATAATGATGATGAACGCAGTTTATT
AATGCTCAAAAGCGTACAACTGCCCTTTGACATCATGGTTTTGGATAACGTCTCCG
```

View signatures on contig NC_002506 | View signatures on contig NC_002505 |

1503 of 55385 SIGNATURES FOR VIBRIO CHOLERAE O1 BIOVAR ELTOR STR. N16961

Sig Start	Sig Stop	Sequence	Gene Start	Gene Stop	Gene Function	Origin	□
564444	564760	TACGTGGTTATGGCAACCTAATCATCATAAAAC ATAATGAGCACTATTTAAGTGCCTATGCCCACA ATGATCAGCTGCTAGCAAAGGAAGGACAAACTG TGCAAGCAGGCCAGAAGATCGCCACCATGGGAA GTTCTGGTACCAACAGCGTACGCTTGCACTTTG AGATCCGTTATCAAGGGAAGTCAGTGAATCCTA AACGCTACTTACCTTAACTCATATTGCTAAGGT CATACTGCAAAGGTTGTAAATCGCAATGACTCG CAAAGTTGCCAGGGGGAGGCCGCTATGAGTGTC AGCAATACCGTAACCAAAGT	563722	564658	lipoprotein NlpD	NC_002505	□
754608	754878	AGTATTCAATACCATCAGCGACAGAAAGCGTTT GATCATTTGTGAATACGATACCCATTCGTTACA AAAATTGGTCACGATCCGACAAATAACCTAATA GCGTAAAAGCCAATATTTTGGCTACGCTGACAG CCTGGATCACCTAACTCTATGTATAATTTGGTC GCATTATAATCAGAGCGGGTGAAGCTGTCAGCT TAAAAGGTTCAATTTATTGACGCTAAATGAATT TCACCTGAGTGGGTTTGACGTTTTCACTAGGGT AGCAGCC	754827	756003	chorismate mutase/prephenate dehydratase	NC_002505	□
221827	222095	TTTTGCGGTAGTGATGCGTGAAGCCCCTTCATT	221337	221970	nucleoid occlusion p	NC_002505	□

Signature Filters

Gene Name:

Gene Function:
e.g. hypothetical protein.

Contain subsequence:
e.g AGGC

Minimum GC Conent: 30%

Minimum Melting Temp (T_m): 45°C

Signature chain length: 99 500

Show only signatures between region: -1 bp ≤ bp

Gene Location:
☑ in genes ☑ intergenic □ No G at 5' end

☑ Show exact-match 20-mers

☑ Show corresponding gene info:
☑ Sort by sequence length:

Figure 3.2: Insignia search results page, showing the graphical genome browser and table of signature sequences. In the genome browser, forward-strand genes are shown in red, reverse-strand genes in blue, and signatures in green. The top of the signature table is visible at the bottom of the page, showing start and end positions, signature sequences, and annotations. Dynamic signature filters are available in the right margin.

The JavaScript genome browser in the center of the page supports dynamic selection, zooming, and panning, and enables users to browse the selected signatures in the context of their surrounding annotation. In this view, forward-strand genes are shown in red, reverse-strand genes are shown in blue, and the signatures are shown in green. Selecting a signature in the genome browser displays its sequence and causes its table row to be highlighted in yellow, as shown at the bottom of the page. Selecting a gene displays its functional annotation and sequence. Further selecting the name or ID of the gene, links to the GenBank page describing the gene product.

To provide responsive interaction, the signature table and genome browser are limited to displaying at most 2,000 signatures. Often, many more signatures are returned from the search and must be filtered to reduce the number. The easiest solution is to gradually increase the length filter until less than 2,000 signatures remain. Adjusting any of the filters on the right of the page dynamically updates the signatures displayed in the genome browser and signature table. Signatures can be filtered by gene name, gene function, sequence composition, GC content, melting temperature, length, and location. For example, clicking “Show corresponding gene info” and entering “toxin” in the “Gene Function” filter shows only genes with the word “toxin” in their functional annotation. Performing this filter on the results of the *Vibrio cholerae* query given on the Examples page returns a number of long signatures contained in a gene annotated as “toxin secretion ATP-binding protein.” With this functional filter, assays designers can target known virulence genes necessary for pathogenicity. Such signatures would be capable of detecting a virulent gene in any genomic context (e.g. within a genetically engineered bacterium).

After filtering the signatures to obtain a manageable number of candidates, users can search the signatures with BLAST and design PCR primers to target the signature. Users can select “Run BLAST Search“ from the Submit menu, to upload the selected signatures to the NCBI BLAST website to perform a more sensitive search and view any alignments to the background. Once the uniqueness is confirmed by BLAST, users can select “Design Primers” from the Submit menu to design suitable primers for the region using an integrated version of Primer3 [27]. Primer3 can be run with default parameters, or with user-defined constraints for assay-specific experimental conditions. TaqMan assays, with a single probe between the primers, can also be designed with the integrated Primer3 software.

3.2.3 Implementation

The Insignia software is organized into two primary components: the Web interface and the computational pipeline. The computational pipeline is a standalone component, written in C++, that generates match data offline on a computational cluster and stores the results in indexed files. Signature queries can be transmitted to the pipeline, which retrieves and processes match data from the database files to compute signatures. The dynamic Web interface submits queries to the computational pipeline and displays the results. The Web interface is written in HTML, JavaScript, PHP, and Perl. The signature, sequence, and annotation data used by the Web interface are stored in a mySQL database using a Chado schema [59].

3.3 Summary

Insignia is a transparent, highly accessible signature pipeline and database that can be queried for signatures from a Web interface, making it accessible to users without access to high-performance computing resources. The Insignia Web application can indentify an exhaustive set of accurate DNA signatures for any set of target genomes—screening these signatures against a comprehensive background of over 100 billion nucleotides that includes all sequenced bacteria and viruses, the human genome, and many other animals and plants. To date, thousands of Insignia searches have been performed via the Web interface.

3.4 Author Contributions

A version of this chapter appeared previously in published form:

Phillippy AM, Ayanbule K, Edwards NJ, Salzberg SL (2009) Insignia: a DNA signature search Web server for diagnostic assay development. Nucleic Acids Res 37: W229-234.

I designed and implemented all computational methods, engineered a prototype Insignia interface, and wrote the text. Kunmi Ayanbule engineered the current Insignia interface and managed the Insignia mySQL database. Nathan Edwards provided the *e*-unique signatures and drafted their description. Steven Salzberg guided the project and edited the text.

Chapter 4

Pan-genome Tiling Arrays[†]

DNA signatures are capable of rapidly identifying a pathogen via PCR-based assays, but cannot provide a comprehensive picture of the detected genome. Alternatively, array comparative genomic hybridization is a fast and cost-effective method for detecting, genotyping, and comparing the genomic sequence of unknown bacterial isolates. This method, as with all microarray applications, requires adequate coverage of probes targeting the regions of interest. An unbiased tiling of probes across the entire length of the genome is the most flexible design approach. However, such a whole-genome tiling requires that the genome sequence is known in advance. For the accurate analysis of uncharacterized bacteria, an array must query a fully representative set of sequences from the species' pan-genome. Prior microarrays have included only a single strain per array or the conserved sequences of gene families. These arrays omit potentially important genes and sequence variants from the pan-genome.

[†] This chapter includes previously published work with multiple authors. See Section 4.7 for details.

This chapter presents a new probe selection algorithm (*PanArray*) that can tile multiple whole genomes using a minimal number of probes. Unlike arrays built on clustered gene families, PanArray uses an unbiased, probe-centric approach that does not rely on annotations, gene clustering, or multi-alignments. Instead, probes are evenly tiled across all sequences of the pan-genome at a consistent level of coverage. To minimize the required number of probes, probes conserved across multiple strains in the pan-genome are selected first, and additional probes are used only where necessary to span polymorphic regions of the genome. The viability of the algorithm is demonstrated by array designs for seven different bacterial pan-genomes and, in particular, the design of a 385,000 probe array that fully tiles the genomes of 20 different *Listeria monocytogenes* strains with overlapping probes at greater than twofold coverage.

4.1 Background

Microarrays are well known for their success in studying gene expression [60]. As one of their many other roles, DNA microarrays can also be used to characterize both large-scale and small-scale genetic variations. For instance, array comparative genomic hybridization (CGH) is commonly used in human cancer studies to genotype cell lines by detecting gene loss and copy number variations [61]. At a finer resolution, microarrays are also used to detect single nucleotide polymorphisms at targeted loci [62]. In addition to human screens, microarrays have been widely used for the detection and genotyping of microbial species. Notably, a viral genotyping microarray [9] was one of the methods used to etiologically link severe acute respiratory syndrome (SARS) to a novel coronavirus [63]. Arrays for the detection and comparative analysis of bacterial genomes have also been developed, including arrays for *Listeria monocytogenes* [64,65,66,67,68], and many other bacterial species. However, these earlier, low-density arrays did not contain enough probes to target the entire genome of the bacterium, and were forced to probe only a small subset of the known genes.

As the density of DNA microarrays increased in recent years, it has become possible to probe the entire genome of an organism in addition to only specific genes. An array providing unbiased coverage of probes across a genome is commonly referred to as a whole-genome tiling array. Such arrays have been very successful for genome-scale analysis, including the discovery of novel transcripts, splicing variants, protein binding sites, and polymorphisms [69]. Depending on the offset between adjacent probe locations, whole-genome tilings can be either gapped, end-to-end, or overlapping Figure 4.1a.

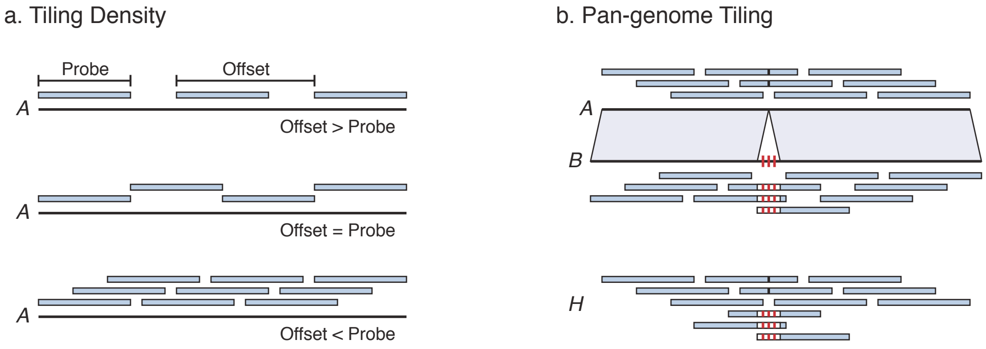


Figure 4.1: Illustration of tiling array densities and pan-genome tiling. Genomes are represented as horizontal lines and probes as colored rectangles. The *offset* between probes is the distance between the start of one probe and the start of the next. a) Three different tiling densities are shown for genome *A*. The top figure illustrates a gapped tiling, the middle an end-to-end tiling, and the bottom an overlapping tiling. b) A pan-genome tiling is shown for two genomes. Genomes *A* and *B* are identical except for a small insertion in *B*, represented by vertical red bars. Solid blue probes are conserved in both genomes, and probes spanning the insertion event are colored by variant. Set *H* shows the non-redundant set of probes needed to tile the pan-genome including *A* and *B*.

In the human genome, tiling arrays are designed to probe the genome at evenly spaced intervals. To maximize the expected specificity of the array, repetitive probes must be avoided and experimental conditions, such as melting temperature, equalized. This creates an optimization problem in choosing which sequences should be included on the array [70,71]. In smaller microbial genomes, it is possible to target every position of the genome with overlapping probes, simplifying the design process. For example, extreme high-density arrays can now accommodate 2.1 million variable length probes on a single chip (Roche NimbleGen, Inc). For an average 2 Mb sized bacterial genome and 50 nt probe length, probes can be offset by only a single base pair and still span the entire genome, generating coverage of 50x. By tiling the entire genome, some suboptimal probes will be included on the array, but can be identified and corrected for in the analysis. These overlapping arrays are capable of identifying polymorphism at a much finer resolution than gapped arrays.

Tiling arrays have traditionally been constructed based on the genome of a single reference strain and used to locate genomic differences contained in the experimental strains. However, single-genome arrays can only detect and analyze sequences similar to those included on the array, and cannot discover or analyze sequences absent in the reference strain. After the introduction of the pan-genome concept [72,73], it has become increasingly clear that some microbial species contain significant genetic diversity, and it is not suitable to compare against only a single reference strain. The pan-genome hypothesis states that any given species has two sets of genes. First, a set of core genes present in all strains that define the species; and second, a set of dispensable genes present in only one or a few of the strains that presumably mediate adaptation. A single genome describes the genomic material for a particular strain, but the pan-genome describes the genomic makeup for an entire species. Single reference tiling arrays cannot survey this full diversity. Ideally, an array for analyzing new strains should cover the genomic diversity of the entire pan-genome.

With the explosion in microarray densities, it is now possible to design pan-genome tiling arrays that contain all genomic sequence from the known pan-genome. The simplest strategy is to fully tile the genomes of each strain independently. However, due to similarities between the strains, some sequences would be tiled with excessive redundancy, and this approach would be cost ineffective. Instead, a pan-genome array should aim to minimize costs by using the minimal probe set necessary to target every element of the pan-genome with adequate coverage. The typical approach for targeting multiple strains is to group individual genes into gene families and then probe only the conserved sequences of those families [74,75,76]. For example, Willenbrock *et al.* designed an innovative 32 strain *Escherichia coli* pan-genome array by clustering homologous genes based on pairwise alignment similarity [74]. Homology was defined as gene alignments with an E-value $< 10^{-5}$, a bitscore > 55, and alignment coverage of at least 50% of the gene length. For each resulting gene group, a consensus sequence was generated via multiple alignment, and probes were designed to target the most conserved regions of the consensus. The resulting array comprised 224,805 probes, targeting 9,252 gene groups, with a median coverage of 27 probes per gene group.

Targeting only the conserved sequence of gene families is an effective and efficient method for detecting—at a low resolution—the presence and absence of gene families; however, for studies that require a finer resolution, this method omits many potentially significant sequences from the array. Firstly, a slight variation in a gene (e.g. a partial deletion) can be responsible for a significantly different phenotype. By only targeting the conserved portion of gene families, the variable regions responsible for these differences will not be included on the array. Secondly, a gene-centric design includes only coding sequences. Therefore, these designs cannot be used to detect differences in intergenic regions, which may include regulatory elements, or used for studies that require a whole-genome tiling, such as transcriptome mapping or chromatin-immunoprecipitation-chip (ChIP-chip) studies. Finally, gene-centric design models depend on an accurate annotation of the genome. If genes have been mis-annotated or omitted from the annotation, such genes cannot be properly represented on the array. This is particularly troublesome for many draft-quality genomes that have highly fragmented sequence assemblies and lack accurate annotations. For these reasons, a whole-genome tiling is preferable for applications that require more flexibility or an unbiased tiling of the genome. However, no methods have been described for efficiently tiling multiple whole-genome sequences.

This chapter describes a method for pan-genome tiling array design that both minimizes the number of probes required and guarantees that all sequences in the pan-genome are fully tiled by the array. The prior gene-centric approaches are abandoned in favor of a more concrete, probe-centric approach that relies only on the genomic sequences and not the annotation. To summarize the new approach, let the pan-genome *G* be the set of all genomes from a species, and let *P* be the non-redundant set of all length *k* substrings from *G*. Due to sequence conservation between genomes, a single probe may match to multiple locations (genomes) of the pan-genome. Call these matches the probe targets. The *Pan-Tiling* problem is to find a minimum cardinality subset $H \subseteq P$ such that all sequences of *G* are targeted by probes in *H* and no target is offset more than *maxoff* from the preceding target or sequence start.

Constructing a full tiling of the pan-genome seems like it would require a large number of probes, but by leveraging the similarities between strains, a reasonably sized probe set can be constructed that fully covers a large pan-genome with adequate redundancy. The key to the strategy is choosing probes that will hybridize to as many of the strains as possible, while using only a necessary amount of probes to cover polymorphisms (insertions, deletions, variants). For example, Figure 4.1b shows a pan-genome tiling for two miniature genomes, with a *maxoff* of one-third the probe length. Genomes *A* and *B* are identical except for a small insertion in the middle of *B*. Fully tiling both genomes requires a total of 19 probe targets (9 for *A* and 10 for *B*), but probe set *H* illustrates that these 19 targets can be tiled with just 12 probes. Conserved probes are used to tile the left and right of both genomes, and distinct probes are used to tile the two polymorphism variants. This is obviously a simplified example. The problem becomes more difficult as the number of genomes and complexity of polymorphisms increases.

The methods presented in this chapter were developed to aid the design of a pan-genome CGH tiling array for *Listeria monocytogenes*—the causative agent of listeriosis and a NIAID category B biodefense agent that is of significant food safety and public health concern [77]. The species of *L. monocytogenes* is composed of three primary genetic lineages (named I, II, and III) that display different capabilities of environmental survival and pathogenic potential to cause human infectious disease [78]. In order to both characterize new strains based on genetic content, and detect polymorphism at a higher resolution in small RNAs (sRNAs) and intergenic sequences, the array was required to cover all pan-genomic sequences with a high density of probes. This bacterial species is particularly well suited for pan-genome array design because there are a remarkable number of strains that have been sequenced. At the time of array design, a total of 20 *L. monocytogenes* complete or draft genome sequences were available, totaling 57.9 Mbp (Table 4.1). Genomic sequences and annotations were obtained from The National Microbial Pathogen Database Resource (NMPDR) [79]. The sequence conservation for the sequenced strains was computed with Nucmer [80], and ranges between 92% and 99% in nucleotide identity versus the completed EGD-e reference strain. Even with such substantial diversity within the species, the PanArray algorithm is able to design a pan-genome tiling covering each genome at more than twofold coverage using only 385,000 50-mer probes. A similar density tiling for a single *L. monocytogenes* strain would require 125,000 probes, meaning the PanArray design covers 20x more genomes using only 3x more probes. A description of this design, along with array designs for six other bacterial pan-genomes, is presented after the methods.

Table 4.1: *Listeria monocytogenes* sequences included on the array.

Strain	Lineage	Serotype	Bases	Contigs	Genes[2]
EGD-e	II	1/2a	2,944,528	1	3,002
LO28	II	1/2c	2,910,810	529	5,078
FSL F2-515	II	1/2a	2,586,267	1,415	NA
FSL J2-003	II	1/2a	2,878,206	406	4,686
1/2a F6854	II	1/2a	2,950,285	133	3,028
FSL N3-165	II	1/2a	2,886,689	33	2,963
J2818	II	1/2a	2,971,223	38	3,270
F6900	II	1/2a	2,958,319	35	3,333
J0161	II	1/2a	3,051,828	51	3,252
10403S	II	1/2a	2,866,709	32	2,944
FSL J2-064	I	1/2b	2,899,431	327	3,914
4b H7858	I	4b	2,972,254	181	3,187
FSL J1-175	I	1/2b	2,902,346	357	4,559
FSL N1-017[1]	I	4b	2,857,865	77	3,465
HPB2262	I	4b	3,006,068	75	3,319
FSL J1-194	I	1/2b	2,986,227	44	3,792
4b F2365	I	4b	2,905,187	1	2,987
FSL R2-503	I	1/2b	3,001,696	54	4,863
FSL J2-071[1]	IIIA	4c	3,149,923	46	3,789
FSL J1-208	IIIB	4a	2,260,760	1,494	NA

All sequences and annotations were obtained from NMPDR.
[1] These genomes were later identified as being mislabeled.
[2] Number of annotated protein coding genes and RNAs reported by NMPDR at the time of the array design.

4.2 Array Design Algorithm

The general strategy of the PanArray design algorithm is best summarized by analogy to the well-known *Minimum Hitting Set* problem in computer science [81,82]. Let P be a set of n points and $F = \{P_1, P_2, \ldots, P_m\}$ be a family of m subsets of P. *Minimum Hitting Set* is the problem of selecting the minimum cardinality subset $H \subseteq P$ such that H contains at least one element from each subset in F. Although finding a minimum hitting set is known to be NP hard, it is a well studied problem and efficient approximation algorithms are known.

To see the similarities between the *Pan-Tiling* and *Minimum Hitting Set* problems, let the sequence G be a concatenation of all the genomes from a species, and let $W = \{w_1, w_2, \ldots, w_m\}$ be the set of m intervals that results from segmenting G into non-overlapping, end-to-end, length l windows. Let P be the non-redundant set of length k substrings from G. A probe candidate $p \in P$ is said to hit a window $w \in W$ if a match between p and a substring of G begins in the interval w. Let $P_i \subseteq P$ be the subset of probes that hit the window w_i, and $F = \{P_1, P_2, \ldots, P_m\}$ for the m windows of W. A minimum hitting set H of F is a minimum cardinality subset of probes $H \subseteq P$ such that every window of the pan-genome is hit by at least one probe in H. Therefore, finding H effectively tiles the entire pan-genome using a small number of probes. This forms the inspiration for the PanArray algorithm.

4.2.1 Window and Probe Indexing

Windowing the genome simplifies the *Pan-Tiling* problem by casting it terms of the familiar *Minimum Hitting Set* problem, and at the same time enforces the *maxoff* constraint. Because each window is forced to contain at least one target, any two adjacent targets cannot be separated by more than twice the window length. Therefore, the window length is equal to one half *maxoff*. For example, given a maximum offset of $2l$, windows are marked off every l bases of the pan-genome—with the first window w_1 covering the interval $[1, l]$, and the second window w_2 covering $[l+1, 2l]$, and so on. Assuming one target is chosen per window, and the target locations are evenly distributed within windows, the average distance between adjacent targets is expected to be equal to the window length. For a window length l, equal to the probe length k, the resulting depth of coverage averages one, because the probes are spaced k bases apart on average. For any other window length l, the resulting depth of coverage c is expected to be $c \approx k / l$. The extreme case being $l = 1$, which results in exactly k-fold coverage because a probe must hit every position in G.

To identify a hitting set, once the pan-genome is discretized into a set of windows, each window must be mapped to the set of probe hits it contains. As before, a probe p hits a window if a match between p and G begins within the window's interval. Thus far exact matches have been assumed, but a match can be defined by any criteria necessary for efficient hybridization. To help reduce probe redundancy, the PanArray implementation can optionally use inexact matches containing a single mismatch. Any suitable k-mer indexing algorithm can be utilized for this phase, but allowing for mismatches can be computationally expensive. The implementation uses a fast, but memory intensive, compressed keyword tree for indexing all probe hits. Alternatively, a slower, but memory efficient, hashing scheme would also work. To index the 1-mismatch hits, each probe's $3k$ possible 1-mismatch permutations are added to the index as well. The result of the indexing is a list of positions and windows for all k-mers of the pan-genome (the probe candidates). At this stage, the final list of probe candidates may be manually filtered based on typical criteria such as melting temperature, GC content, secondary structure, etc. For ungapped tilings, it is impossible to avoid suboptimal probes. However, highly repetitive probes can be identified by the number of genomic positions they map to, and should be discarded if they threaten to confound the array analysis (e.g. by affecting normalization). Alternatively, the input sequences may be masked prior to k-mer indexing to avoid repetitive or unwanted sequence altogether.

For CGH arrays, each probe is considered equivalent to its reverse complement, but for expression or transcriptome arrays, forward and reverse strand probes must be considered independently. Probe matches are listed on the strand on which they appear, so for single-stranded samples, the sequence to be synthesized for the array will need to be reversed complemented. For DNA tiling arrays it is helpful to assume the sample will be double-stranded so that genomic inversions in one or more of the strains do not have to be tiled separately.

4.2.2 Probe Selection

As described above, finding the minimum hitting set of *P* effectively tiles the entire pan-genome using a small number of probes. As before, *W* is the windowed pan-genome. Let W_p be the subset of windows hit by probe *p*, and *U* be the set of currently uncovered windows. Let a window hit by at least one probe be termed as covered, and the coverage of a probe be the number of windows it hits $|W_p|$. A naive algorithm for finding a small hitting set *H* is to choose, for each uncovered window, a probe hitting the window that also hits the most other windows. The idea being that choosing probes with the highest coverage will minimize the total number of probes necessary to cover all windows. However, this approach does not properly account for the probe coverages. Only a single probe is needed to cover a window, so after selecting a probe *p*, all other probes that hit a window in W_p will see their effective coverage reduced. Take for instance two probes *p* and *q* that hit the exact same set of windows. Choosing *p* reduces the effective coverage of *q* to zero, because all of *q*'s windows have already been covered by *p*. Let the residual coverage r_p of a probe be the effective coverage after some other set of probes have already been chosen ($r_p = |W_p \cap U|$).

A greedy algorithm first suggested by Johnson [83] improves on the naive approach by allowing to reconsider the residual coverage of probes after each iteration. This algorithm has since been shown to be essentially a best-possible approximation for the *Minimum Hitting Set* problem [84]. When adapted for the current problem, the algorithm chooses, while uncovered windows remain, the probe that hits the most currently uncovered windows. The Greedy PanArray Algorithm is:

```
Greedy PanArray Algorithm
    $H = \emptyset$
    $U = W$
    while $U \neq \emptyset$
        select $\operatorname{argmax}_{p \in P} |W_p \cap U|$
            $U \leftarrow U - W_p$
            $H \leftarrow H \cup \{p\}$
    return $H$
```

The algorithm itself is straightforward, but it must be carefully implemented to run efficiently. It is infeasible to recompute the residual coverage $|W_p \cap U|$ for all W_p during each iteration, because both P and W can be on the order of millions for a large pan-genome. To avoid this complexity, the PanArray implementation exploits a property of the residual coverages that allows it to recompute only a few values at each iteration. Note that for any p, its residual coverage r_p can never increase. A probe's coverage either remains the same, or decreases because one of its windows was hit by the prior iteration. Therefore, instead of recomputing all residuals after each iteration, it is sufficient to maintain a priority queue of residual coverages and only update stale values at the front of the queue.

At the start of the algorithm, all initial coverages are inserted into the queue. To maintain the priority queue after a new probe is chosen, all residual coverages are considered invalid. During the next iteration, a new r_p value is computed for the front of the queue, marked as valid, and reinserted into the queue. This process is repeated until a valid residual returns to the front of the queue. Often, newly computed residuals will return quickly to the head of the queue before the others have been updated. At this point it is unnecessary to update any other residuals because their new values cannot be greater than their current value. Therefore, the head of the queue must be the updated maximum. This lazy evaluation of the residuals avoids many unnecessary computations and drastically improves the performance of the algorithm. The greedy algorithm without this speedup takes days to complete, but with the speedup runs in a matter of seconds.

4.2.3 Probe Annotation

The flexibility of the PanArray design algorithm is a result of its probe-centric approach. Because it does not require any identification or clustering of genes, the design is independent of any genome annotation. Therefore, instead of building the annotation into the design of the array, the annotation can be mapped onto the array after the design. Most importantly, this strategy allows for intergenic sequence and unannotated genomes to be included on the array, and annotation updates to be incorporated as they become available. For example, after the *L. monocytogenes* array had been designed, over 40 new sRNAs were discovered in *Listeria* [85].

Conveniently, the sequences of each had already been tiled by the array design, and the updated annotation was easily remapped onto the array. As another example, the gene counts provided by NMPDR in Table 4.1 are inconsistent and vary between 3,000 and 5,000 genes per genome, suggesting considerable annotation error. Uncoupling the array design from the annotations removes any possibility that annotation errors will affect the design.

Included with the final probe set *H* is the list of locations on the pan-genome that each probe matches. If the genome sequence is updated, the location information can be easily recovered by remapping the probes to the genome using a matching tool such as MUMmer [80] or Vmatch [26]. To annotate the array, probes are mapped to all annotation features with a coinciding location. The result is a many-to-many mapping with each feature being targeted by multiple probes, and a single probe possibly targeting multiple features (e.g. conserved genes between strains). With this mapping, all probes targeting a specific gene in the pan-genome can be quickly recovered.

4.2.4 Implementation

The PanArray algorithm was implemented in C++, and the source code is freely available at (http://www.cbcb.umd.edu/software/panarray). The Listeria monocytogenes array design described above is available from the Gene Expression Omnibus [86] under GEO accession number GPL8942.

4.3 Listeria monocytogenes Pan-genome Array

As suggested earlier, *L. monocytogenes* is a good candidate for constructing a pan-genome tiling array because the species has been widely sequenced, with 20 complete or draft genome sequences available. To confirm that the sequenced genomes contain the majority of *L. monocytogenes* genetic diversity, the pan-genome size was estimated using the methods of Tettelin *et al.* [73] as implemented in the Ergatis package (http://ergatis.sourceforge.net). Seventeen of the eighteen *L. monocytogenes* genomes listed as annotated by NMPDR in Table 4.1 were used in the analysis (strain 1/2a F6854 was unavailable at the time). According to the cited method, the addition of an N^{th} genome was simulated by searching the annotated genes of each genome against all possible permutations of N–1 other genomes. Genes without a match over 50% protein similarity for at least 50% of their length were recorded as "new". The number of new genes n expected to be discovered in the N^{th} sequenced genome was modeled by the power law $n = \kappa N^{-\alpha}$, and the parameters κ and α were estimated from the data via non-linear least squares regression on the means using the R function nls [87]. A power law model was found to fit the *L. monocytogenes* data better than the originally proposed exponential model. This agrees with a recent suggestion that a power law is a more appropriate model of the pan-genome phenomenon [88].

The estimated number of undiscovered genes is shown in Figure 4.2. The power law exponent α was found to be 1.38±0.002, suggesting that the *L. monocytogenes* pan-genome is closed (i.e. has a finite number of genes), and the sequencing of more genomes would eventually sample the entire set of dispensable genes. Therefore, it appears the vast majority of *L. monocytogenes* genes have been sequenced and are included on the array. This model predicts that the addition of a 21st genome would yield less than 7 new genes. However, only a single lineage III genome was included in this analysis, so this prediction might be artificially low for a new lineage III strain. The sole lineage III strain analyzed (FSL J2-071) contains 31 genes absent in any of the lineage I and II strains.

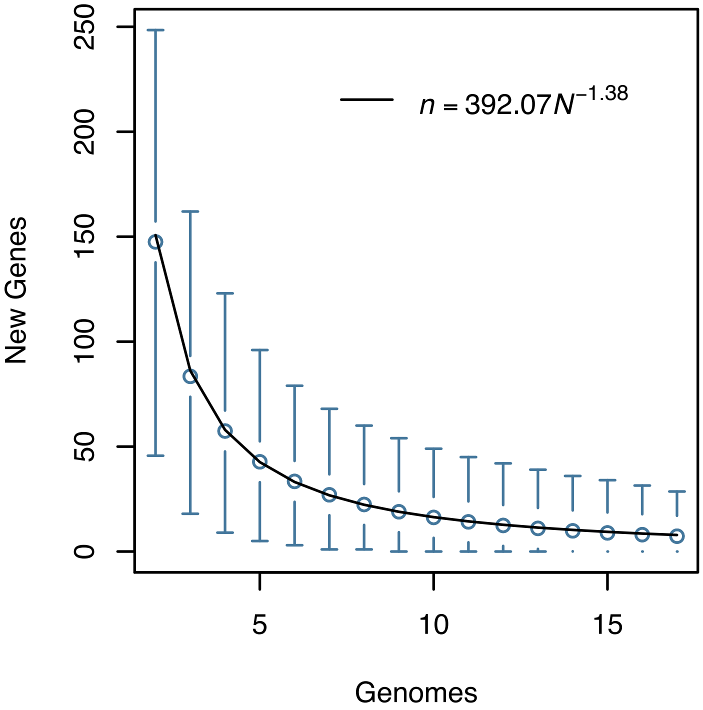


Figure 4.2: *Listeria monocytogenes* new genes regression. The number of new genes *n* predicted to be discovered with the addition of an N^{th} *Listeria monocytogenes* genome sequence. A power law fit to the simulated data is given by the solid curve. The circles represent the mean value for each *N*, and error bars show the 90% confidence intervals.

To capture the full diversity of *L. monocytogenes*, all 20 genomes listed in Table 4.1 were included in the design, with a combined sequence length of 57,946,621 bp and a total of 65,431 annotated genes. To avoid tiling low quality or contaminant sequence, contigs less than 2 Kbp in length were discarded—reducing the tiled sequence length to 54,810,759 bp. The design was constrained to a 385,000 feature NimbleGen array with a probe length of 50 nt. Because hybridization of a 50-mer probe will tolerate a few mismatches, probes differing by a single mismatch were considered equivalent during the design phase. The window length was set to 24 bp, enforcing a maximum target offset of 48, an expected depth of coverage of about 50 / 24 = 2.08x, and resulting in approximately 2.3 million windows. These parameters guarantee that every base pair of the pan-genome will be covered by at least one probe, since the maximum offset is less than the probe length.

To cover each window, the PanArray algorithm selected 373,389 distinct probes mapping to 2,893,387 positions in the pan-genome. On average, each probe in the design targets about 8 different positions in the pan-genome. Rather than being repeated sequences within the same genome, these different locations most often refer to a conserved locus in multiple strains Figure 4.3. Interestingly, the degree of probe reuse corresponds well with the known evolutionary relationship of the strains. Included on the chip are 8 genomes from lineage I, 10 from lineage II, and 2 from lineage III. This would suggest that the peak at *Genomes* = 1 in Figure 4.3 is for strain-specific probes; the peaks around 2 and 9 are for lineage-specific probes; and the peak around 20 is for species-specific probes that are conserved in all 20 *L. monocytogenes* genomes.

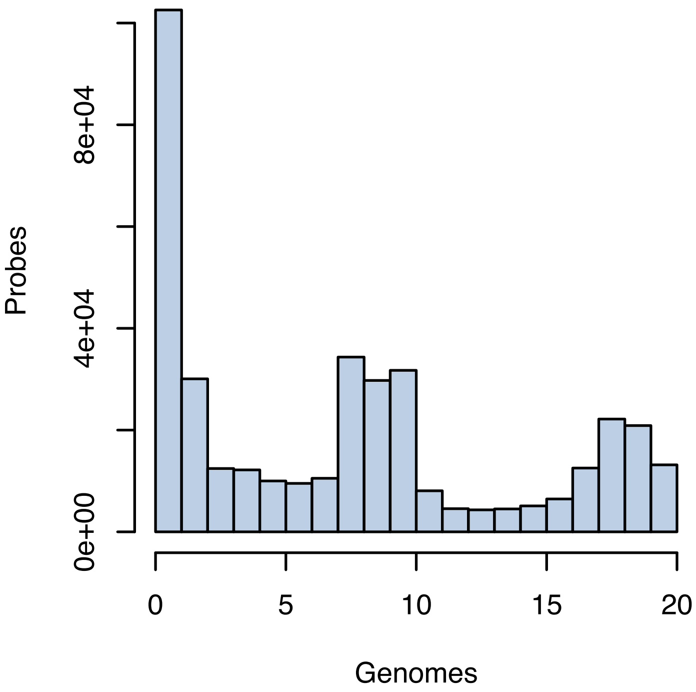


Figure 4.3: PanArray probe reuse histogram for *Listeria monocytogenes*. The number of genomes targeted by each 50-mer probe is given on the horizontal axis. Targets may contain up to one mismatch to the probe.

Because this is a dense tiling of the entire genome, it was unnecessary to optimize probes for uniqueness, as is done in standard expression arrays with only a few probes per gene. Probes were screened for repetitive sequences, but the L. monocytogenes strains were found to contain few repeats. The most repetitive 15-mer occurs only 28 times per genome, and the most repetitive 50-mer probe used in the design targets a “cell wall surface anchor protein” family and occurs a maximum of 16 times per genome. Altogether, 99.2% of the probes target at most one location per genome.

To augment the original PanArray design, an additional 228 negative control probes were added to the array, chosen from *Bacillus* spp., which is a known cohabitant of *Listeria*. The negative control probes were chosen to be specific to

Bacillus spp. using Insignia. The remaining 11,838 features on the array were filled by selecting individual probes to supplement the lowest coverage regions of the design. All probes were checked to conform to NimbleGen design specifications, and a few probes were trimmed to meet synthesis cycle limits. The resulting *L. monocytogenes* pan-genome array has an average depth-of-coverage of 2.65x, with a median probe offset of 21 bp, and a modal offset equal to the window length of 24 bp. The full distribution of probe offsets is given in Figure 4.4. As expected, the average offset is equal to the window length (24 bp). The uneven distribution and pronounced mode is the caused by non-random tie breaking. In the case of a conserved sequence, where every probe hits the same number of genomes, the first probe of the window is always chosen. Also, the heavy left tail indicates that many windows are covered by more than one probe and the solution that is slightly denser than expected (2.65x actual vs. 2.08x expected). This may be a consequence of the sequence composition, or may indicate a non-optimal solution. Finally, the majority of targeted sequences exactly match their probe (75%) and the remainder match with a single mismatch (25%).

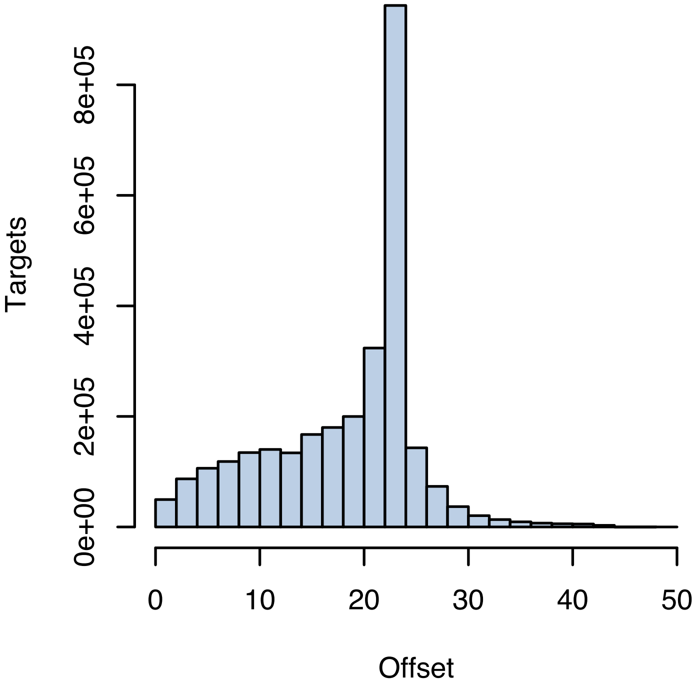


Figure 4.4: PanArray probe offset histogram for *Listeria monocytogenes*. The offset between two adjacent probe targets is given on the horizontal axis. Targets may contain up to one mismatch to the probe.

The performance gain of PanArray over more naive methods is significant. For instance, selecting a single probe from each window requires roughly 2.3 million probes. The slightly more principled naive algorithm, that does not recompute residual coverages, chooses 1,739,242 probes, but is still well over the 385,000 probe limit. The Greedy PanArray algorithm meets this limit and vastly outperforms the other methods—requiring only 373,389 probes to cover the entire pan-genome. With the lazy evaluation speedup, the PanArray algorithm is also comparable in runtime to the naive algorithms. On a single 2.4 GHz processor, the naive algorithm took 29 seconds; the greedy algorithm without lazy evaluation was terminated without completing after a few days; and the Greedy PanArray algorithm with lazy evaluation took only 130 seconds. The runtime for the final design process was dominated by building the *k*-mer index, which required 84 minutes using a custom implemented compressed keyword tree.

4.4 Array Requirements for Bacterial Pan-genomes

Using PanArray, additional arrays were designed for a total of seven bacterial pan-genomes, for which a large number of genomes have been sequenced. The additional species include: *Francisella tularensis*, *Staphylococcus aureus*, *Bacillus anthracis*, *Vibrio cholerae*, *Burkholderia pseudomallei*, *Escherichia coli*, and *Shigella* spp. Due to their high similarity, *E. coli* and *Shigella* spp. were considered as a single pan-genome. To facilitate easy comparison, all designs were created with a window length of 25 bp, a probe length of 50 nt, and allowing for probes to contain a single mismatch to their target. As with the *L. monocytogenes* design above, draft genomes were included, but contigs less than 2 Kbp were discarded. The results are given in Table 4.2. Probe "reuse" is measured in the average number of targets per probe. It is rare for a 50-mer probe to match to more than one location per genome, so the number of targets per probe is roughly equivalent to the average number of genomes that a probe matches.

Table 4.2: Summary of PanArray probe requirements for various species.

Species	Strains	Avg. Length[1] (Mbp)	Pan Length[2] (Mbp)	Targets[3]	Probes	Reuse[4]
F. tularensis	14	1.88	26.29	1,355,504	121,312	11.2 (0.80)
S. aureus	14	2.88	40.38	2,006,144	200,999	10.0 (0.71)
B. anthracis	9	5.48	49.29	2,230,870	246,947	9.0 (0.99)
L. monocytogenes	20	2.74	54.81	2,832,489	358,688	7.9 (0.39)
V. cholerae	15	3.87	58.09	3,017,198	346,447	8.7 (0.58)
B. pseudomallei	20	6.72	134.31	6,755,234	491,231	13.8 (0.69)
E. coli / Shigella	29	4.96	143.72	8,210,679	674,697	12.2 (0.42)

[1] The average genome length for a species.
[2] The sum of all genome lengths for a species.
[3] The total number of locations targeted by the probes. A single probe may target multiple genomes in the species.
[4] The average number of targets per probe. In parentheses, the reuse divided by the number of genomes.

The highly conserved species of *B. anthracis* exhibits near perfect probe reuse. Almost every *B. anthracis* probe matches all of the included strains; therefore, the number of probes required to tile the nine sequenced strains is nearly the same as is required to tile one strain. This is because the pan-genome of *B. anthracis* is closed and the strains are highly conserved at the nucleotide level (usually containing only a few SNPs per strain). Adding successive *B. anthracis* strains to the array would increase the required number of probes very gradually.

In contrast, *L. monocytogenes* has the lowest degree of probe reuse, with each probe targeting on average only 39% of the included strains. This is a reflection of the diversity of strains that have been sequenced and the low level of nucleotide conservation between strains, with some strains averaging a single nucleotide polymorphism rate as high as 8%. Any SNP rate of higher than 2% (1 per 50 bp) exceeds the 1 mismatch threshold per probe and requires additional probes to target the divergent sequence. However, as more variants are added to the array, the addition of each successive genome requires fewer new probes than the last, on average. Figure 4.5 shows this relationship for the *L. monocytogenes* strains. Successive strains are added by order of lineage, from the bottom of Table 4.1 to the top, and the design is recomputed at each step. There are pronounced jumps in the number of probes required when the first of a new lineage is added, but the number of probes needed to tile the rest of the lineage quickly levels off.

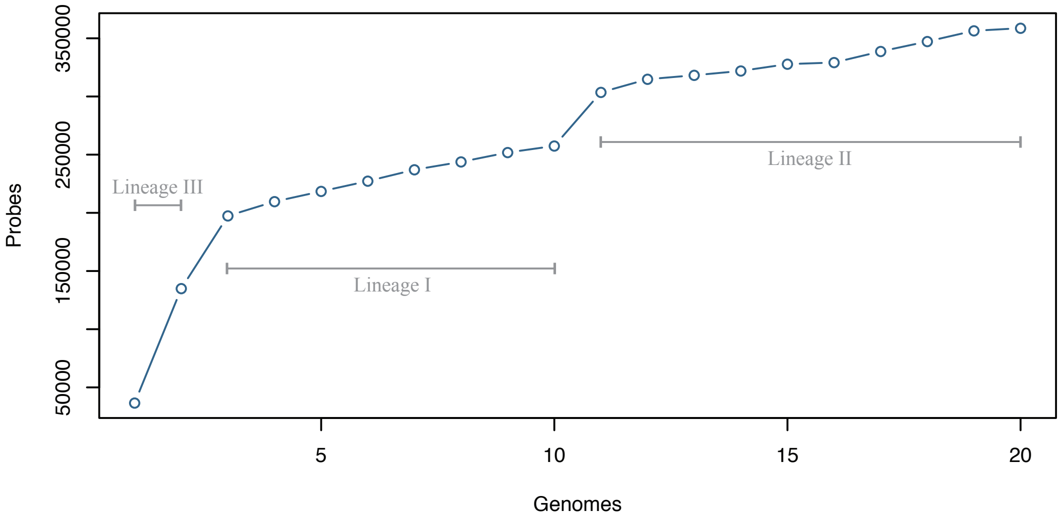


Figure 4.5: PanArray probe requirements for *Listeria monocytogenes*. The number of probes required by PanArray to tile the *L. monocytogenes* pan-genome with the successive addition of each genome. Genomes are added by order of lineage and the design recomputed after each addition.

Escherichia coli and *Shigella* spp. form the largest pan-genome currently sequenced, totaling over 144 Mbp of genomic sequence. Even for a pan-genome of this size and diversity, PanArray effectively tiles all sequences at an average of 2x coverage using only 674,697 probes—well below the maximum number of probes available on current arrays. The *B. pseudomallei* pan-genome is roughly equivalent in total number of pan-genome bases, but requires considerably fewer probes because of higher probe reuse. Due to the large number of sequenced genomes and relatively high similarity between strains, the *B. pseudomallei* design exhibits the highest probe reuse factor of all the designs (13.8x). Creating a 2x coverage tiling by choosing one probe every 25 bp would require roughly 5.4 million probes for the *B. pseudomallei* pan-genome, but PanArray was able to create a 2.5x tiling of the same pan-genome with only 491,231 probes.

4.5 Discussion

The PanArray algorithm described above is ideal for high-density tilings of overlapping or closely spaced probes. The results have shown that this algorithm is applicable for all currently available bacterial pan-genomes. However, if the maximum number of probes is limited, or the genome size is extremely large, it may be necessary to design a tiling with gaps between the probe targets (i.e. a maximum offset greater than the probe length). In this case, it is necessary to choose unique probes that avoid unwanted cross hybridization between repetitive sequences within the genome. To achieve this, repetitive probes can be filtered, or the coverage scores used in the PanArray algorithm can be weighted to penalize repetitive probes. For example, probe coverage can be redefined as the number of genomes a probe targets, rather than the number of windows, and probes targeting multiple windows in the same genome can be appropriately down-weighted. In many cases, probes within the same window will share the same coverage score, and rules can be applied for breaking the tie and choosing the most reliable probe. Similar schemes could be devised to favor probes with any other desirable criteria.

Array analysis of CGH experiments is typically conducted on signal ratios between a reference and experimental hybridization. Duplications or deletions in the experimental samples are evident as non-zero values of the log ratio of the two normalized signals. So-called segmentation algorithms examine this log ratio across multiple positions in reference sequence to determine the boundaries of the variations [89,90]. The most accurate methods consider not just individual probes, but a context of probes around a genomic location, and can identify even small polymorphisms between the strains. These analyses require both a reference signal and a reference coordinate system on which the probes are tiled. Usually a whole-genome tiling is constructed for a single reference strain, but because PanArray provides a whole-genome tiling for every reference strain included in the array, the same array design can be used to perform segmentation analysis against any reference strain on the array.

In addition to segmentation analysis versus a reference genome, a pan-genome array makes it possible to analyze uncharacterized strains in the context of the entire pan-genome. In some cases, it is preferable to use a multi-strain control [91], but depending on the number of genomes, it can be impractical to co-hybridize all reference strains included on the chip. In these cases, traditional segmentation or log-ratio analysis must be replaced by a method that does not require a reference hybridization signal. For gene-level analysis, direct analysis of the individual probe intensities provides comparable sensitivity and specificity versus segmentation analysis [74], and various methods have been developed that operate independently of a signal ratio [74,92,93]. A probe-based approach provides the most flexibility for pan-genome array analysis, because each probe can be individually scored based on its own intensity, and the genes can be classified based on the aggregated scores of the individual probe scores without the need for a control hybridization. A new analysis method designed in this spirit is described in the next chapter.

Pan-genome tiling arrays have all the applications of single-strain tiling arrays, but with enhanced flexibility and the ability to analyze previously uncharacterized strains. Pan-genome CGH offers an economical alterative to sequencing for determining the genomic makeup of uncharacterized strains in a species and explaining the causative factors of phenotypic differences between strains. Probe based methods, like microarray, are especially well suited for situations where sequencing is inefficient because there is a low abundance of target DNA and a high abundance of background DNA intermixed. For example, applications such as real-time pathogen detection, surveillance, and diagnostics require a known sequence of DNA to be targeted from a vast environment [19,24,94]. A pan-genome array could be used for the detection and genotyping of pathogens from a large environment, without needing to isolate the individual cells. Pan-genome arrays could also be used to capture all species- or locus-specific genomic material from an environment, which could then be directly processed or sequenced separately from the metagenome. Microarray based genomic capture has already been applied to targeted human resequencing as an efficient means of enriching for desired sequencing templates [95,96,97].

4.6 Summary

Without the need for sequencing additional genomes of the same species, pan-genomic CGH has become an increasingly popular and cost-effective approach to compare and characterize genomic contents of unknown bacterial isolates. Prior multi-strain arrays have targeted the conserved sequences of gene families, or a selected group of polymorphisms; therefore, providing only partial coverage of the pan-genome. PanArray is a probe selection algorithm capable designing a tiling array that fully covers all genomes of a species using a minimal number of probes. The viability of this method is demonstrated by array designs for seven different bacterial pan-genomes, each of which can fit on a single microarray slide. By constructing an unbiased tiling of all known sequences, these unique pan-genome tiling arrays provide maximum flexibility for the analysis, detection, or capture of genomic material for entire species.

4.7 Author Contributions

A version of this chapter appeared previously in published form:

Phillippy AM, Deng X, Zhang W, Salzberg SL (2009) Efficient oligonucleotide probe selection for pan-genomic tiling arrays. BMC Bioinformatics 10: 293.

I designed and implemented all computational methods, performed the analysis, and wrote the text. Wei Zhang, Xiangyu Deng, and Steven Salzberg motivated the problem and contributed edits to the text.

Chapter 5

Probing the Pan-genome of *Listeria monocytogenes*[†]

Listeria monocytogenes is a foodborne bacterial pathogen widely known for its adaptability to diverse habitats and host niches, and its high fatality rate among infected, immunocompromised individuals. At least three major genetic lineages have been defined within this species. Although all three lineages are considered to be pathogenic to human, it is intriguing that only two lineages (LI and LII) account for the vast majority of sporadic and epidemic cases of human listeriosis, whereas the third lineage (LIII) is rarely implicated in human infections for unknown reasons. This chapter probes the genomic diversity of 26 *L. monocytogenes* strains by employing the unique pan-genomic DNA array described in the previous chapter. Comparative genomic hybridization of 9 lineage III strains on this array uncovered 86 protein-coding genes and 8 small regulatory RNAs that are highly specific to the predominant listeriosis-associated lineages, which potentially contribute to bacterial stress adaptation, host niche fitness and pathogenicity. Exponential regression analysis predicts that *L. monocytogenes* has a core genome of between 2,330 to 2,456

[†] This chapter includes work in preparation with multiple authors. See Section 5.10 for details.

genes (80% of each individual genome) and a pan-genome repertoire of over 4,052 unique genes. Comparison of all lineage strains reveals high genomic synteny with moderate sequence drift associated with lysogenic bacteriophages. Phylogenomic reconstructions based on 3,560 homologous groups suggest a polyphyletic population infrastructure and gradual loss of metabolic genes as this saprophytic species diversified into the rare and probably defective lineage III.

5.1 Background

Listeria monocytogenes is a Gram-positive foodborne bacterial pathogen and the causative agent of the human and animal infectious disease, listeriosis. *L. monocytogenes* can thrive in diverse environmental reservoirs (e.g. soil, water, and sewage) and proliferate under unfavorable conditions (e.g. high osmolarity, low pH, and refrigeration temperature) that other bacterial pathogens cannot endure [98,99,100,101]. Its robust physiological characteristics, coupled with its ubiquity in food processing, distribution, and retail environments, have made *L. monocytogenes* difficult to manage in food manufacturing, particularly for ready-to-eat food products. *L. monocytogenes* causes the highest rates of hospitalization (about 92%) and mortality (about 20%) among all foodborne bacterial pathogens in the United States [102], making the control of this bacterium in foods a high priority for both food safety and public health. Yet, the versatile lifestyle of *L. monocytogenes* both inside and outside its host, and its unique capability to invade and replicate in different host cell types (e.g. macrophages and nonprofessional phagocytes), have made this opportunistic pathogen a paradigm for studying host-pathogen interactions, pathophysiology, gene regulation, and stress adaptation [103,104].

Previous molecular subtyping studies have collectively suggested that the species of *L. monocytogenes* is composed of at least three major evolutionary or genetic lineages that notably differ in their prevalence in causing human and animal diseases [65,78,105,106,107,108,109,110]. Specifically, lineage I (or LI) and lineage II (or LII) of *L. monocytogenes* are frequently isolated from foods and implicated in the vast majority (>95%) of both sporadic cases and epidemic outbreaks of human listeriosis [100]. Genetic lineage III (or LIII) strains are rarely reported in cases of human infections, but are sometimes associated with animal disease cases [100,109,111]. The mechanisms underlying the biased predominance of certain *L. monocytogenes* genetic lineages in human listeriosis remain largely unknown. Several recent studies have revealed elevated levels of genetic diversity among LIII isolates [107,110]. Multilocus sequence typing analysis on the basis of partial *sigB* and *actA* gene sequences have also suggested that LIII is polyphyletic, with the co-existence of at least three distinct subgroups (i.e. LIIIA, LIIIB, LIIIC) [109]. Atypical phenotypes of LIII isolates, such as deficiency in rhamnose fermentation [109], attenuated virulence potential [111], reduced resistance to heat and cold stresses [112], and lowered biofilm productivity [113], have indicated that LIII may have followed a distinct evolutionary path from other *L. monocytogenes* lineages.

Compared to fairly extensive studies on LI and LII strains, little is known about LIII. Although it is documented that most listerial virulence factors such as the positive regulatory factor (or PrfA) are well conserved across the entire *L. monocytogenes* species, LIII strains are underrepresented in both food contamination and human listeriosis. This suggests the existence of additional, yet-to-be-identified genetic factors in the predominant disease-causing *L. monocytogenes* lineages LI and LII that may mediate listerial niche adaptation, resistance to extra- and intracellular stresses, and pathogenicity. These unknown genetic factors may have been lost, mutated, or decayed in LIII as the genomes evolved, resulting in a defective phenotype for LIII isolates in certain ecological and host niches. To test this hypothesis, this chapter combines *in silico* comparative genomic analyses with an array-based comparative genomic hybridization (CGH) approach to probe the genomic diversity of *L. monocytogenes* and to identify genomic features common in LI and LII but absent in LIII. Array CGH is a powerful yet cost-effective approach for genotyping and detecting intraspecies genomic diversity for many bacteria. Previous efforts on comparative genomic analyses underscore the usefulness of CGH in resolving genetic lineages and identifying strain- or lineage-specific genes in *L. monocytogenes* [64,65,66,67,68]. However, most of these studies targeted only a number of selected genes or partial listerial genomes, making an accurate assessment of intraspecies genomic diversity difficult.

It is recognized that a few sequenced genomes may not fully represent the entire genetic repertoire of a given organism [72,73,88,114,115]. For this reason, the pan-genome concept has triggered new investigations on genomic diversity for several bacterial species, including *Streptococcus* spp. [73,116,117], *Haemophilus influenzae* [118], *Neisseria meningitides* [119], *Escherichia coli* [74,120,121], and *Lactococcus lactis* [122]. Pan-genome refers to the total genetic repertoire of a given species, which is typically composed of "core" genes plus some "dispensable" or "accessory" genes [115,123]. Pan-genomic DNA arrays, that probe this full repertoire, have recently gained increasing popularity for the systematic survey of diversity in prokaryotic species [74,124,125].

The availability of more than 20 sequenced *L. monocytogenes* full and draft genomes has made this pathogen an ideal candidate for pan-genomic study (Table 5.1). Initial comparative analysis of 17 *L. monocytogenes* genomes in Chapter 4 indicated a "closed" pan-genome for this bacterial species. Species with a closed pan-genome typically share highly syntenic genomes with less frequent horizontal gene transfers and genomic rearrangements. Therefore, the entire gene pool can be fully sampled by sequencing a small set of representative isolates, and the number of new genes to be discovered by sequencing additional genomes will quickly approach zero. This prompted the design of a pan-genome CGH array that, in theory, accommodates the total genomic diversity of the *L. monocytogenes* species on a single DNA chip. Compared to several previous pan-genome microarrays that targeted either the conserved sequence of gene families with low probe density or no coverage of the intergenic regions, PanArray designed a pan-genome tiling array that incorporates the full genomes of 20 available *L. monocytogenes* strains [126], providing unbiased coverage of the pan-genome and superior accuracy and resolution for data analysis.

Using integrated data obtained from both *in silico* whole-genome comparisons and pan-genome CGH analyses, this chapter aims to (1) explore the intraspecific genetic diversity of *L. monocytogenes* with a focus on the largely unexplored genetic lineage III; (2) estimate the core and pan-genome that define the *L. monocytogenes* species; (3) identify unique protein-coding genes and regulatory RNAs in the predominant disease-causing lineages, as they may relate to niche adaptation and pathogenicity; and (4) reconstruct an accurate phylogeny for different *L. monocytogenes* lineages and strains based on pan-genome characteristics.

Table 5.1: Comparatively analyzed *L. monocytogenes* genomes.

Strain	Lineage	Serotype	Size (bp)	Contigs[1]	Genes[2]	% Identity[3]	Note[4]
EGD-e	II	1/2a	2,944,528	Closed	2931	100	Array design, CGH
R2-561	II	1/2c	2,945,851	37	2993	99.78	Array design
LO28	II	1/2c	2,675,580	1150	3030	99.6	Array design
Finland 1988	II	3a	2,834,040	49	2740	98.49	Data analysis
10403S	II	1/2a	2,873,541	21	2905	98.48	Array design
F2-515	II	1/2a	1,815,995	1728	2710	98.47	Array design
N3-165	II	1/2a	2,884,080	39	2885	98.39	Array design
J2-003	II	1/2a	2,741,640	795	2972	98.32	Array design
F6900	II	1/2a	2,968,620	23	3007	98.28	Array design
F6854	II	1/2a	2,950,285	133	2967	98.26	Array design
J2818	II	1/2a	2,973,040	24	3020	98.24	Array design
J0161	II	1/2a	3,062,582	25	3114	98.23	Array design
J1-175	I	1/2b	2,866,484	457	3178	94.39	Array design
J2-064	I	1/2b	2,828,700	545	2968	94.37	Array design
R2-503	I	1/2b	2,991,493	55	2968	94.28	Data analysis
J1-194	I	1/2b	2,989,818	30	3040	94.27	Array design
N1-017	I	4b	3,142,060	79	3253	94.2	Array design[5]
Clip 80459	I	4b	2,912,690	Closed	2972	94.17	Data analysis
F2365	I	4b	2,905,187	Closed	2907	94.14	Array design
H7858	I	4b	2,972,254	181	3195	94.08	Array design
HPB2262	I	4b	2,991,120	79	3067	93.98	Array design
HCC23	III	4a	2,976,212	Closed	3059	92.38	Data analysis
F2-524	IIIA	4a	-	-	-	-	CGH
F2-501	IIIA	4b	-	-	-	-	CGH
J2-071	IIIA	4c	2,851,800	53	2778	92.6	Array design, CGH[5]
J1-208	IIIB	4a	1,963,740	1660	2809	91.8	Array design, CGH
M1-002	IIIB	4b	-	-	-	-	CGH
W1-111	IIIB	4c	-	-	-	-	CGH
F2-208	IIIC	4a	-	-	-	-	CGH
F2-569	IIIC	4b	-	-	-	-	CGH
W1-110	IIIC	4c	-	-	-	-	CGH

[1] Number of contigs based on GenBank at the time of the study; strains with >200 contigs were excluded from analyses due to high risk of sequencing, assembly and annotation errors.
[2] Number of annotated protein coding genes and RNAs based on GenBank.
[3] Nucleotide sequence identity in reference to strain EGD-e as computed by Nucmer.
[4] Strains used for array design; comparative genomic hybridizations; data analysis in this study.
[5] Strains N1-017 and J2-071 were found to be mislabeled in GenBank; this has since been fixed.
-Information not available.

5.2 Pan-genomic Array Completeness

Initial power-law regression analysis of 17 fully sequenced *L. monocytogenes* genomes suggested that this bacterial species exhibits a nearly closed pan-genome, which would yield rapidly diminishing returns of less than 7 novel genes per additional genome sequenced. Therefore, a single array could presumably query the full genetic repertoire of the species, and be used to completely genotype currently unsequenced strains. For this purpose a pan-genomic array comprising 385,000 50-mer *in situ* synthesized oligonucleotide probes was designed that fully tiles the sequences of 20 *L. monocytogenes* genomes, with no gaps, at greater than 2-fold coverage of each genome.

Shortly after completion of the chip design, four additional *L. monocytogenes* genomes were sequenced to closure, including strain Clip 80459 (LI), strain Finland 1988 (LII), strain R2-561 (LII) and strain HCC23 (LIII). These new *L. monocytogenes* genomes enabled evaluation of the genomic coverage of the PanArray design by individually mapping each of the 385,000 oligonucleotide probes to annotated genes in the four genomes. A 50-mer probe was mapped to a particular gene if it perfectly matched the gene sequence or contained only a single nucleotide mismatch. For each annotated gene, the probe coverage was calculated as the percentage of the gene length covered by mapped probes Table 5.2. These results suggest that the array adequately represents the intra-species diversity of *L. monocytogenes*, particularly for LI and LII genomes. However, due to the limited number of fully sequenced LIII genomes available at the time of design, the coverage for LIII specific genes is less optimal, as indicated by HCC23.

Table 5.2: Probe coverage for newly sequenced genomes.

		Probe coverage		
Genome	Lineage	100%	90%	80%
R2-561	II	0.95	0.98	0.98
Clip 80459	I	0.91	0.99	0.99
Finland 1988	I	0.80	0.96	0.98
HCC23	III	0.30	0.80	0.89

Proportion of genes from four newly sequenced strains with probe coverage meeting a minimum percentage of the gene length (100%, 90%, 80%) for probes containing at most one SNP.

5.3 Accuracy of the Array

Genomic DNAs for nine LIII strains were each co-hybridized on the pan-genomic arrays with that of EGD-e as an internal reference. The nine LIII strains were carefully selected from a strain collection to represent 3 different serotypes (4a, 4b, and 3c) as well as 3 different subgroups (IIIA, IIIB, and IIIC) of *L. monocytogenes* LIII. Individual probes were designated as present or absent in the sample based on statistical analysis of the normalized signal intensities (see section 5.7.2 Pan-genomic Array Analysis). Since the position of each probe is known for all sequenced *L. monocytogenes* genomes, genes were scored by the fraction of targeting probes with a positive signal, otherwise known as the positive fraction (PF). This yields a very flexible scoring scheme that can be readily applied to any intragenic or intergenic feature of the genome targeted by a sufficient number of probes. A high PF indicates a gene is likely present in the hybridized genome. Circular maps of all PF values for the nine LIII genomes in reference to a LI strain F2365 and a LII strain EGD-e are shown in Figure 5.1.

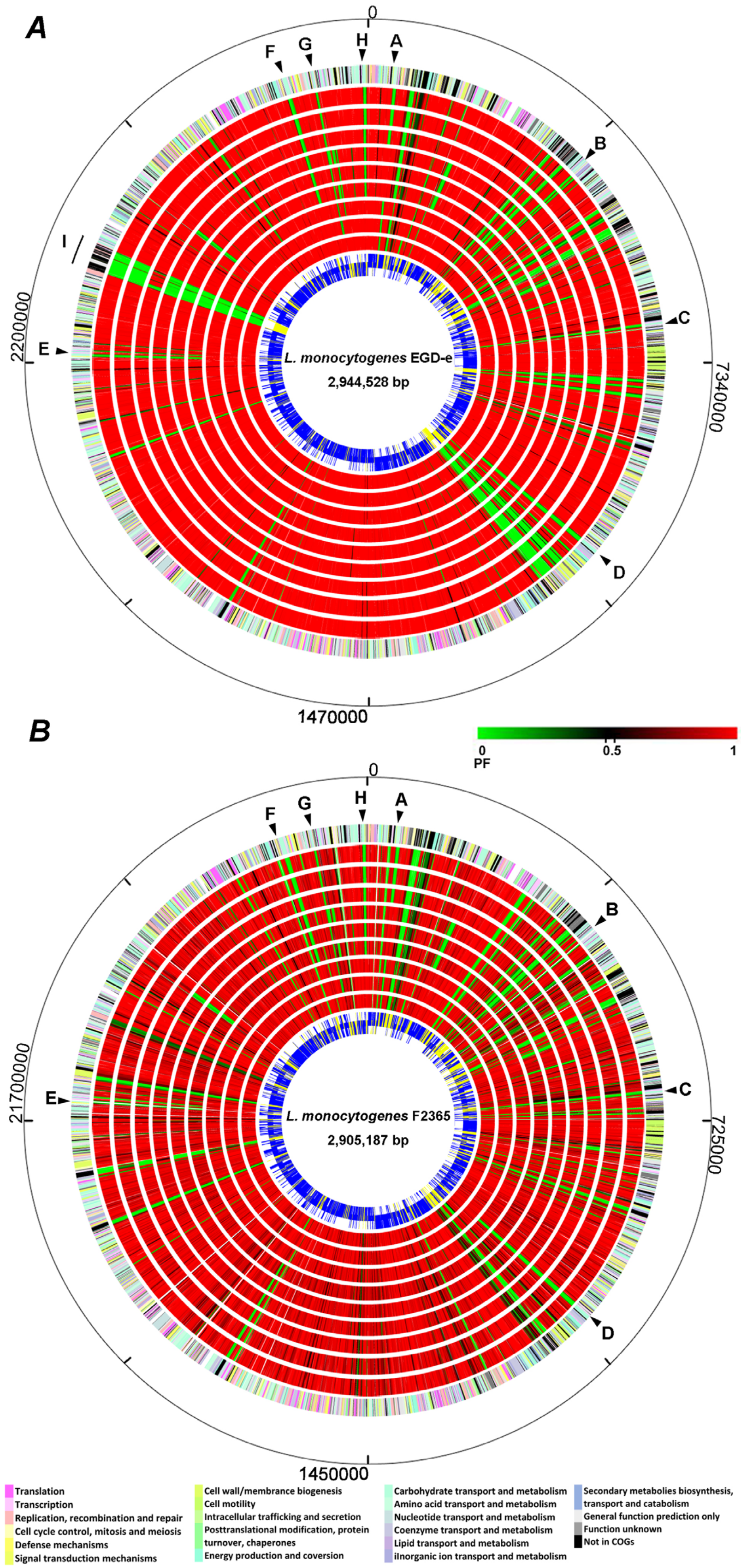

A
0
F
G
H
A
B
C
I
E
2200000
734000
D
1470000
L. monocytogenes EGD-e
2,944,528 bp
B
0
0.5
1
PF
F
G
H
A
B
C
E
2170000
725000
D
1450000
L. monocytogenes F2365
2,905,187 bp
Translation
Transcription
Replication, recombination and repair
Cell cycle control, mitosis and meiosis
Defense mechanisms
Signal transduction mechanisms
Cell wall/membrance biogenesis
Cell motility
Intracellular trafficking and secretion
Posttranslational modification, protein turnover, chaperones
Energy production and coversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Nucleotide transport and metabolism
Coenzyme transport and metabolism
Lipid transport and metabolism
iInorganic ion transport and metabolism
Secondary metabolies biosynthesis, transport and catabolism
General function prediction only
Function unknown
Not in COGs

Figure 5.1: *Listeria monocytogenes* circular gene maps compare the genomes of nine LIII strains with that of a LII reference strain EGD-e (*A*) and a LI reference strain F2365 (*B*). The inner most circle is the reference genome. Core genes in the reference genome are shown blue and accessory genes are shown in yellow. From inside out, the second to the tenth circles represent the nine LIII genomes, including J2-071 (LIIIA), F2-501 (LIIIA), F2-504 (LIIIA), J1-208 (LIIIB), M1-002 (LIIIB), W1-111 (LIIIB), F2-208 (LIIIC), F2-569 (LIIIC), and W1-110 (LIIIC), respectively. Genes in LIII genomes are color-coded based on the PF values (see the reference bar). Green indicates a gene is absent (PF=0) in a LIII genome; red indicates a gene is conserved (PF=1) in a LIII genome at the corresponding location in the reference genome. The eleventh circle gives color-coded gene annotations in the reference genome based Clusters of Orthologous Groups of proteins (see the color codes at the bottom). The outer most circle provides relative genomic coordinates. Eight DDG clusters associated with carbohydrate transport and metabolism and absent in LIII strains at similar genomic locations in EGD-e and F2365 are marked with letters A through H. Specifically: A, *lmo0037-0041* (or *lmof2365_0045-0050*); B, *lmo0357-0360* (or *lmof2365_0377-0381*); C, *lmo0631-0633* (or *lmof2365_0660-0662*); D, *lmo1030-1036* (or *lmof2365_1051-1057*); E, *lmo2133-2138*; F, *lmo2732-2736* (or *lmof2365_2719-2723*); G, *lmo2771-2773* (or *lmof2365_2761-2763*); and H, *lmo2846-2851* (or *lmof2365_2836-2841*), respectively. The LII-specific *comK* prophage integration region was marked in the EGD-e genome (I).

To select an appropriate PF threshold and test the accuracy of gene calls based on PF values, true-positive and false-positive rates were computed for the PF criterion on 51,814 annotated *L. monocytogenes* genes, compared against genomes for which there was both sequence and microarray data. True gene "presence" was determined by a tblastn search of the 51,814 predicted proteins against a six frame translation of the genome [31], requiring a minimum of 50% amino acid similarity and an E-value $\leq 10^{-5}$. Figure 5.2 shows the ROC curves for the PF criterion measured against the tblastn standard for two *L. monocytogenes* strains, EGD-e and J2-071. The PF measure is remarkably robust, as there appear to be very few genes near the classification threshold. For example, Figure 5.3 shows a density estimation of PF values for both present and absent genes, showing that the vast majority of present genes have PF > 0.9 and absent genes PF < 0.1. Based on the ROC analysis, a PF cutoff of 0.6 was chosen to best match the tblastn results and minimize the expected error rate. The seemingly higher false-positive rate for J2-071, in comparison to the closed EGD-e genome, is partially due to tblastn false-negatives incurred from the 78 gaps in the J2-071 draft genome. In these cases, a gene that is truly present, but overlapping a sequencing gap, is falsely reported as absent by the tblastn method, which artificially increases the measured false-positive rate of the CGH array method.

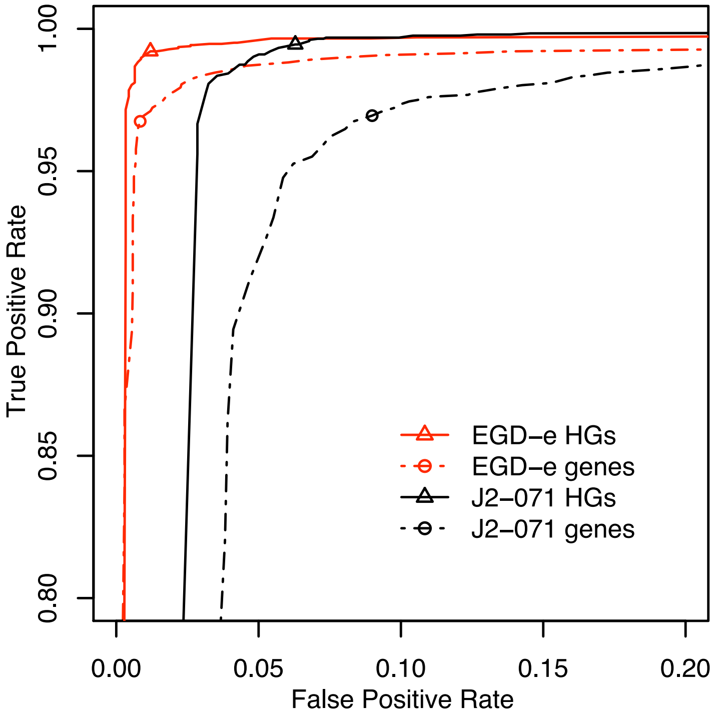


Figure 5.2: PanArray receiver operating characteristic curves. ROC curves compare true-positive rates with false-positive rates of different PF cutoffs for prediction of the presence or absence of individual gene variants and homologous groups. Error rates are shown for genes (dotted lines) and homologous groups (solid lines), computed from EGD-e (red) and J2-071 (black) control hybridizations. Circles indicate the chosen PF cutoff of 0.6 for classifying gene variants. Triangles indicate the chosen PF cutoff of 0.6 for classifying homologous groups.

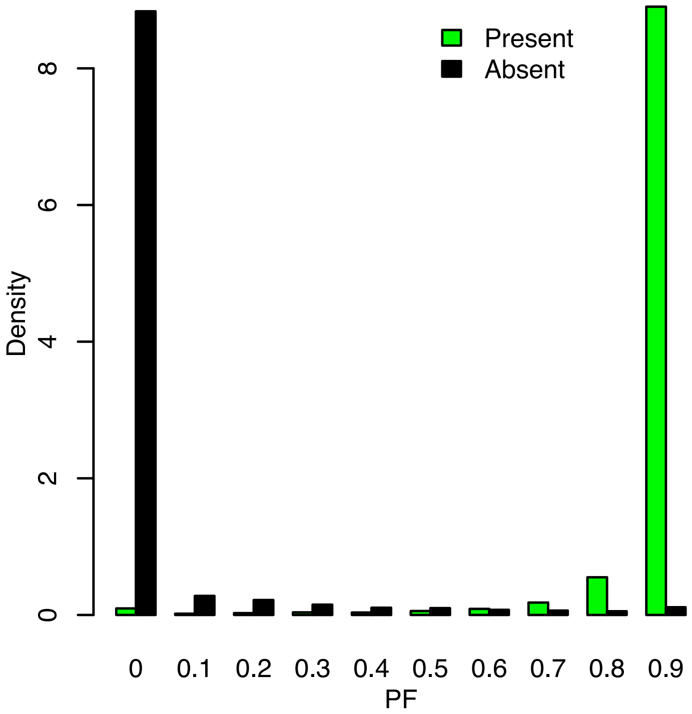


Figure 5.3: PanArray positive fraction probability densities for known present and absent genes, demonstrating the vast majority of truly present genes have PF score greater than 0.9 and the vast majority of truly absent genes have PF less than 0.1. Green bars show the density of PF scores for genes found present by a tblastn search, and black bars show the density of PF scores for genes found absent by a tblastn search. PF labels give the minimum of each left-closed interval. For example, PF=0.5 bars show the densities for the bucket PF=[0.5,0.6).

Accuracy statistics for the chosen 0.6 PF cutoff versus the 50% alignment similarity cutoff are given in Table 5.3. They array has perfect sensitivity for detecting the EGD-e and J2-071 control genes. Accuracy was estimated for detecting both individual gene variants from all other strains and for detecting homologous gene groups (HGs). Orthologous gene groups are typically preferred; however, the inability of CGH to accurately determine sequence identity and gene order makes it impractical to discriminate between highly similar paralogs. Alternatively, 3,560 strongly homologous gene groups, identified by clustering proteins with higher than 50% amino acid similarity, were tested for presence or absence. A gene group was marked as present in a genome if any gene from that group exceeded the BLAST or PF threshold. Figure 5.2 displays the true- and false-positive rates of homologous group detection alongside the original ROC curves. In comparison to detecting individual gene variants, HG detection significantly increases the sensitivity of the array without increasing the false-positive rate, significantly increasing the area under the ROC curve. When analyzing only a single gene variant on the chip, high polymorphism in the sample genome can disrupt hybridization and lead to false-negatives. However, by considering an entire gene group, a sample need only hybridize with its nearest variant, thereby increasing the sensitivity [122]. To demonstrate the sensitivity of the array at detecting HGs in unsequenced strains, Table 5.3 lists accuracy statistics for EGD-e and J2-071 when the probes specific to those genomes are removed from the analysis. This simulates the accuracy of the array at calling genes in an unsequenced LII and LIII strain. The sensitivity of the array is only slightly affected, with a 0.2% true-positive rate drop for EGD-e and a 1.3% drop for J2-071. The drop is more pronounced for J2-071 because it is one of only two linage III genomes included on the array, so ignoring the J2-071 specific probes more dramatically affects the sensitivity of calling HGs from that lineage.

Table 5.3: PanArray accuracy for detecting genes and homologous groups.

Chip Data	Test Data	Present	Absent	ACC[1]	TPR[2]	FPR[3]	FDR[4]
EGD-e	EGD-e genes only	2846	0	1.000±0.000	1.000±0.000	N/A	N/A
EGD-e	All gene variants	49068	2746	0.973±0.002	0.973±0.003	0.020±0.009	0.001±0.000
EGD-e	Gene groups	2642	918	0.989±0.002	0.993±0.001	0.024±0.007	0.008±0.003
EGD-e(-)	Gene groups	2627	918	0.987±0.002	0.991±0.001	0.024±0.007	0.008±0.003
J2-071	J2-071 genes only	2694	0	1.000	1.000	N/A	N/A
J2-071	All gene variants	47411	4403	0.964	0.970	0.090	0.009
J2-071	Gene groups	2543	1017	0.978	0.995	0.063	0.025
J2-071(-)	Gene groups	2468	1016	0.969	0.982	0.062	0.025

Present/Absent are the number of genes present/absent based on a tblastn search. For EGD-e, the mean of 9 data sets are given, along with their standard deviation to illustrate array reproducibility.
[1]Accuracy. (TP+TN) / (P+N). [2]True-positive rate. TP / P.
[3]False-positive rate. FP / N. [4]False discovery rate. FP / (FP+TP).
(-)Excludes all probes directly targeting the test strain from the analysis to simulate accuracy for an unknown strain.

5.4 Core and Pan-genome Estimates

The expected number of new genes to be discovered by sequencing additional *L. monocytogenes* strains, and the sizes of the core and pan-genomes, were estimated using methods adapted from Tettelin *et al.* [73]. However, frequent gaps and sequencing errors in low-quality genome assemblies were found to cause many missed protein alignments, which affected the core genome estimation. For example, only 683 EGD-e proteins meet the alignment threshold in all 24 draft *L. monocytogenes* genomes, an unreasonably low number. Additionally, fragmented annotations in the low quality genomes artificially inflate the pan-genome size estimate. To avoid these artifacts, only 18 “high quality” *L. monocytogenes* genomes, consisting of fewer than 200 contigs each, were used for the new genes and pan-genome estimation. Array CGH results for the 8 additional LIII genomes were included in the core gene estimate.

5.4.1 Core Genome Estimate

To estimate the *L. monocytogenes* core genome, the number of shared genes was computed for many random permutations of *N* genomes ($1 \leq N \leq 26$), and the mean number of shared genes was computed for each *N*. The number of core genes for the species was estimated by fitting an exponential decay function to the means. For the high-quality sequenced genomes, this analysis yielded an estimated horizontal asymptote of 2,467 ± 7 core genes. However, the sequenced genomes include only two LIII genomes. Repeating the analysis for all 26 genomes, including CGH results for the 8 additional LIII genomes, reduced the estimate by over 100 genes to 2,330 ±

5, emphasizing the importance of a balanced sample of diversity for estimating core genome size. Figure 5.4a displays the result of the 26 genome analysis including a smoothed density plot of the shared gene count distributions, the mean value for each *N*, and the best-fit exponential decay.

Imperfect detection sensitivity due to sequencing gaps makes it impossible to achieve convergence for real data, so an exact core genome cannot be determined. Any non-zero false-positive rate for detecting core genes will artificially shrink the core genome with each additional genome, violating the horizontal asymptote of an exponential decay. This is evident in the almost linearly decreasing means towards the tail of Figure 5.4a. To account for these false-negatives, an additional parameter was introduced to the core genes model that adds a constant number of false-negatives upon the addition of each genome (see section 5.7.3 Pan-genomic Sequence Analysis). The revised model is a much closer fit to the data (residual standard error of 2.98 versus 10.68), accounts for noisy draft and CGH data, and yields an increased core genes estimate of 2,456 ± 4 (Figure 5.4a). This likely represents an upper bound on the core genome size. Considering results from both models, and the uncertainty caused by the draft genomes and CGH data, the core genome of *L. monocytogenes* is estimated to be between 2,350 to 2,450 genes (approximately 80% of a typical *L. monocytogenes* genome).

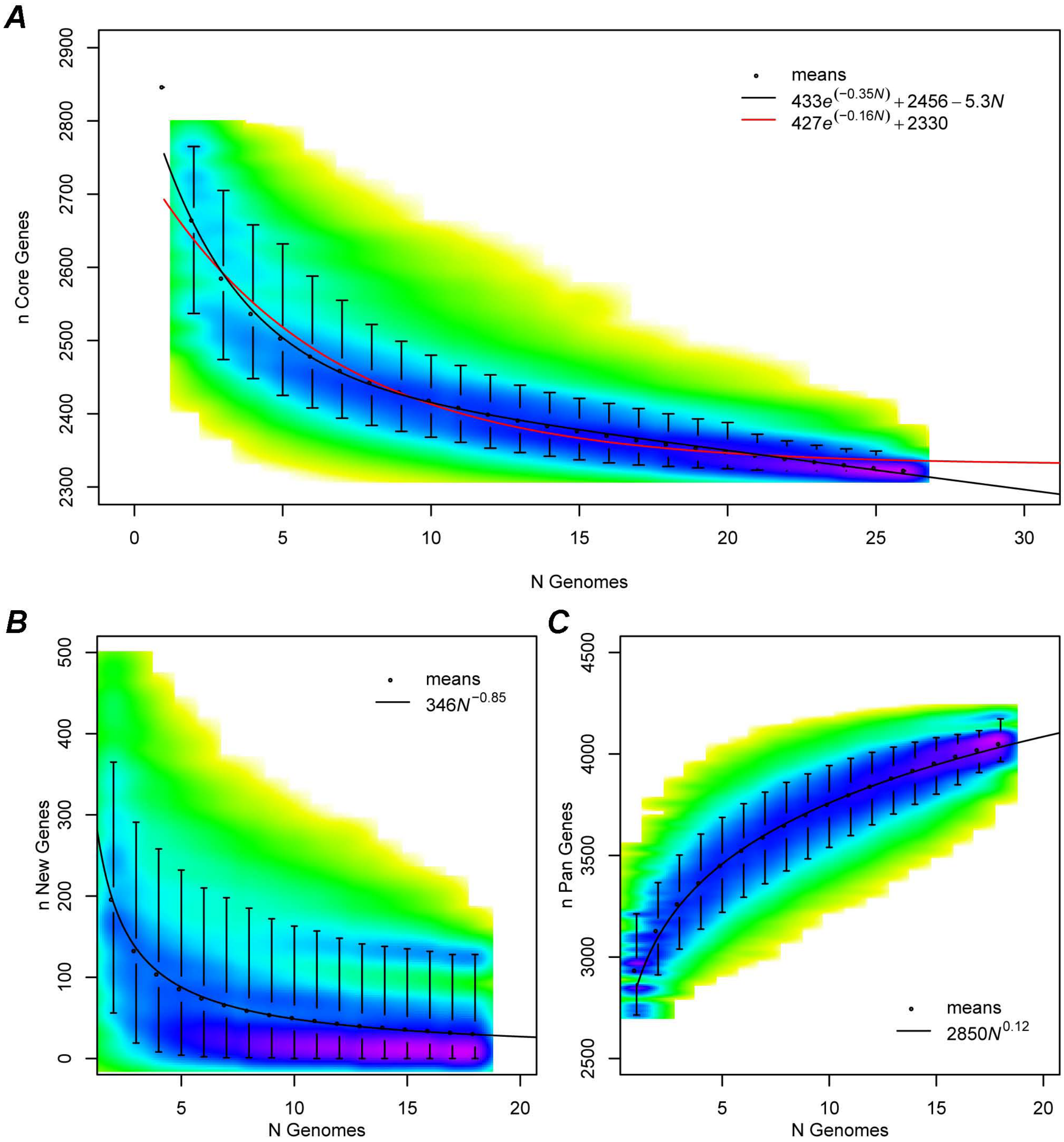


Figure 5.4: *Listeria monocytogenes* core, new, and pan gene regressions. a) Exponential regression analysis that predicts the number of core genes in *N* sequenced genomes. The sampled distribution is represented by a smoothed color density plot obtained through kernel density estimation. Yellow indicates the lowest density and purple indicates the highest density. For each *N*, black circles indicate the mean value and whiskers indicate the 5th and the 95th percentiles of the distribution. An exponential decay fit to the means is given by a solid red curve. A modified exponential decay is given by a solid black curve, which better fits the observed data by accounting for false-negative gene calls. b) Power law regression analysis predicts the number of new genes that will be discovered by sequencing additional *L. monocytogenes* genomes. The LIII genomes are the outliers that pull the means higher, indicating that LIII diversity has not yet been fully sequenced. c) Power law regression analysis predicts the number of *L. monocytogenes* pan genes accumulated from genome sequencing is currently 4,052 and growing with diminishing returns.

5.4.2 Pan-genome Estimate

A major limitation of array CGH is that this method cannot detect novel genes contained in the LIII genomes. For this reason, the pan-genome estimation was performed only for the high-quality sequenced genomes, of which two are from LIII. Again, the number of new genes identified by sequencing each additional genome was computed for many random permutations of N genomes. The number of new genes identified for each N was modeled by the power law function $n = \kappa N^{-\alpha}$ [88]. Using the median values, the power law exponent α was estimated to be 1.12 ± 0.02. This is slightly lower than the previous estimate of 1.38 due to the recent sequencing of four additional genomes, an updated annotation, and a stricter similarity threshold. In both cases, an exponent $\alpha > 1$ indicates a closed pan-genome, meaning the size of the pan-genome is a bounded function of the number of sequenced genomes. However, fitting a power law to the mean values of these distributions yields $\alpha = 0.85 \pm 0.01$, suggesting an open pan-genome (Figure 5.4b). This difference is caused largely by the diverse strains N1-017, HCC23, and J2-071, which contain many strain-specific genes and pull the mean values higher than the medians. For example, strain HCC23 contains 122 strain-specific genes not found in any of the other 17 strains. Removal of these three genomes from the analysis results in an α slightly greater than one for both the mean and median analyses. Two of these strains are the only two high-quality LIII strains available, indicating that additional sequencing of LIII strains may reduce the exponent further and reveal novel and significant diversity in this previously overlooked lineage. This regression analysis suggests *L. monocytogenes* has a significantly diverse gene reservoir, and additional sequencing of LIII genomes is necessary to resolve the exact size and nature of the *L. monocytogenes* pan-genome.

The estimated growth of the *L. monocytogenes* pan-genome with additional sequencing was also simulated using many random permutations of genomes. For open pan-genomes, the cumulative number of unique genes discovered with the sequencing of additional genomes can be modeled by Heap's law using the power law function $n = \kappa N^{\gamma}$ [88]. This regression is illustrated by Figure 5.4c and γ was estimated as 0.12 ± 0.001. Since the growth of an open pan-genome is equivalent to the number of new genes added after sequencing each successive genome, the derivative of the pan genes function should be equal to the new genes function. That is $N^{\gamma-1} \propto N^{-\alpha}$ and $\alpha = 1-\gamma$ for $\alpha < 1$. Although simulated separately, the pan and new gene functions do follow this property for the mean value regressions, with α = 0.85 and γ = 0.12 being in good agreement, and the derivative of the pan genes function nearly equal to the new genes function. For N = 18, the mean estimated pan-genome size is 4,052 and continues to grow, with diminishing returns, for larger N.

This above method is useful for estimating the size of the pan-genome, but because it depends on the order of the genomes analyzed, it does not yield a single representative set of pan genes for the analyzed strains. An alternative that does not depend on the order of genomes is to measure the number of gene groups identified by a similarity clustering method such as OrthoMCL [127]. To be consistent with the other analyses, this method was adapted to cluster strong homologs rather than orthologs. From a graph of 52,776 proteins with >50% similar proteins connected by edges, 3,744 HGs were identified using the MCL graph clustering algorithm [128].

This provides a relative lower bound for the size of the currently sequenced L. monocytogenes pan-genome.

5.5 Lineage-specific and Disparately Distributed Genes

Lineage-specific genes refer to those exclusively present in a single *L. monocytogenes* genetic lineage based on the above defined similarity threshold. Annotated genes in three representative genomes, F2365 (LI), EGD-e (LII), and J2-071 (LIII), were screened for lineage specificity (Table 5.4). To maintain a stringent specificity criterion, a gene was not considered to be lineage-specific if any member from its homologous group was present in other lineages. Using this criterion, only 5 of 21 LII-specific genes previously identified by Doumith *et al.* [129] passed the threshold.

Table 5.4: Lineage specific genes identified in *L. monocytogenes*.

Gene	Genome	Annotation
Lineage I specific		
LMOf2365_0409	F2365	Hypothetical protein
LMOf2365_1251	F2365	Hypothetical protein
LMOf2365_1252	F2365	Hypothetical protein
LMOf2365_2638	F2365	Similar to cell surface anchor family protein
Lineage II specific		
lmo0525	EGD-e	Hypothetical protein
lmo0737	EGD-e	Hypothetical protein
lmo1061	EGD-e	Similar to two-component sensor histidine kinase
lmo1968	EGD-e	Similar to creatinine amidohydrolases
lmo1969	EGD-e	Similar to 2-keto-3-deoxygluconate-6-phosphate aldolase
Lineage III specific		
LmonocytogFSL_030100000415	J2-071	Hypothetical protein
LmonocytogFSL_030100003416	J2-071	Hypothetical protein
LmonocytogFSL_030100004481	J2-071	Hypothetical protein
LmonocytogFSL_030100010091	J2-071	Similar to ADP-ribose 1''-phosphate domain protein
LmonocytogFSL_030100010130	J2-071	Hypothetical protein
LmonocytogFSL_030100011357	J2-071	Hypothetical protein
LmonocytogFSL_030100012027	J2-071	Hypothetical protein

Lineage specificity is based on comparative analysis of 26 genomes in this study, including 7 LI strains (F2365, H7858, Clip 80459, N1-017, R2-503, HPB2262 and J1-194), 9 LII strains (EGD-e, R2-561, Finland 1988, 10403S, N3-165, F6900, F6854, J2818 and J0161) and 10 LIII genomes (HCC23, J2-071, F2-501, F2-524, J1-208, M1-002, W1-111, F2-208, F2-569 and W1-110). Gene ID is designated based on a respective reference genome.

In addition to lineage-specific genes, 86 disparately distributed genes (DDGs) were identified that are highly conserved in the common disease-causing LI and LII strains (PF>0.6) but largely absent or divergent in the rare LIII strains (PF<0.6). DDGs are of particular interest because the biased distribution and conservation of these genes in LI and LII genomes likely correlate to the enhanced ecological fitness and pathogenicity of *L. monocytogenes* in the host. The largest functional group of DDGs (41%) is associated with carbohydrate transport and metabolism. Figure 5.1 illustrates their distribution. *L. monocytogenes* harbors one of the largest bacterial carbohydrate phosphotransferase system (PTS) genes [130,131]. The abundance and diversity of the PTS system allows this soil saprophyte to utilize different carbon sources associated with the ecosystems it inhabits such as soil, silage and sediments. Fifteen PTS genes were identified as DDGs; most are associated with fructose-specific PTS enzyme II components (*lmo0357–0358*, *lmo0631–0633*, *lmo2135–2137*, and *lmo2733*). The distribution of 978 annotated PTS genes and their homologs in all 26 *L. monocytogenes* genomes was surveyed, and 965 (99%) PTS genes were found conserved in all LI and LII genomes and 7 (0.7%) were found specific to LI. In contrast, 137 (14%) PTS genes are absent or divergent in LIII genomes. Diversity in PTS content is most noticeable among the three LIII subgroups, where 48 (4.8%), 137 (14%), and 136 (13.9%) PTS genes are absent in LIIIA, LIIIB and LIIIC, respectively. An interesting distinction among 3 subgroups is that LIIIA strains are capable of fermenting rhamnose, whereas LIIIB and LIIIC strains are deficient in rhamnose utilization [109]. Interestingly, a cluster of six genes (*lmo2846–2851*), which is likely to mediate rhamnose utilization, is missing from all LIIIB and LIIIC genomes. Five genes in this cluster [85] share protein similarities to the rhamnose catabolic pathway in *Escherichia coli* [132,133] and other Gram-positive bacteria such as *Bacillus subtilius*.

The second-largest functional group of DDGs consists of 12 putative transcription factors representing 7 different regulatory gene families. Six are adjacent to PTS genes and possibly involved in regulating carbohydrate metabolism. Four are absent from the non-pathogenic *L. innocua* [130], *L. welshimeri* [134] and *L. seeligeri* [135], suggesting roles in virulence and pathogenicity. One Crp/Fnr (cyclic AMP receptor protein—fumarate and nitrate reduction regulator) family gene *lmo0753* was found to be highly specific to LI and LII but absent in LIII. This Crp/Fnr factor is adjacent to a bile resistance gene *btlB* and shares high amino acid sequence homology to *prfA*.

Multiple DDGs area associated with gastrointestinal (GI) tract adaptation. Two bile-associated genes *btlB* (*lmo0754*) and *pva* (*lmo0446*) are absent in LIII. Both genes help *L. monocytogenes* resist the antimicrobial effects imposed by bile salts during its passage through human GI tract [136]. Loss of these genes lowered tolerance to bile and reduced persistence in murine GI tract [137]. The glutamate decarboxylase (GAD) system mediates the acid resistance in bacteria [138,139,140]. In *L. monocytogenes gadD1* (*lmo0447*) is responsible for growth at mild acidic conditions (pH=5.1) and *gadD2* (*lmo2363*) primarily mediates the resistance to severe acidic stress (pH=2.8) [141]. *gadD2* is conserved in all lineages, whereas *gadD1* and its coupled glutamate: γ-aminobutyrate antiporter *gadT1* (*lmo0448*) are absent in most LIII strains except for J2-071 and HCC23. An arginine deiminase (ADI) system (*lmo0036–0041*) was recently characterized in *L. monocytogenes* [142]. The ADI system plays a role in listerial acid tolerance and may contribute to the enhanced adaptation to acidic conditions in the stomach. It was previously reported that this gene cluster is present in LI and LII but absent from LIII and non-pathogenic *L. innocua* and *L. welshimeri* [142]. The CGH results, however, showed that the ADI gene cluster is also highly conserved in LIIIB. An additional seventeen DDGs have no homolog in the genome of *L. innocua*, including three putative genes encoding LPXTG surface proteins (*lmo0333*, *lmo1666* and *lmo2085*) and *sepA*, a putative virulence factor co-regulated by PrfA and σ^B [143,144].

Complete tiling of the *L. monocytogenes* pan-genome allowed a survey of 100 non-coding small regulatory RNAs with specified 5' and 3' positions [85] in 9 LIII genomes. The majority (87%) of these sRNAs are conserved in LIII genomes, and only eight were found to be absent or divergent in LIII (PF<0.6) (Table 5.5). Noticeably, all eight sRNAs are also absent from *L. innocua*, and five were differentially expressed in intestinal lumen or blood, suggesting roles in host niche adaptation. For example, *ril38* contributes to listerial survival in human blood [85].

Table 5.5: Small regulatory RNAs absent or divergent in LIII genomes.

		Distribution in lineage III [2]								
		IIIA			**IIIB**			**IIIC**		
RNA	**Regulation**[1]	**J2-071**	**F2-501**	**F2-524**	**J1-208**	**W1-111**	**M1-002**	**F2-569**	**F2-208**	**W1-110**
rli62	n/a	-	-	-	-	-	+	-	-	-
rliG	n/a	-	-	-	-	-	-	-	+	-
rli38	↑ in broth & blood	+	-	-	-	-	-	-	-	+
rli48	↑in intestine	-	-	-	-	-	+	-	+	-
rli26	↑in blood	+	+	+	-	-	-	-	-	-
rli29	↑in intestine & blood	-	-	-	+	-	+	+	-	-
rli49	n/a	-	-	-	-	-	-	-	-	-
rliC	↓in blood	+	+	+	-	-	-	-	-	+

[1] Up-regulated "↑", or down-regulated "↓" *in vivo* [85]; n/a, information not available.
[2] Gene is either present "+" or absent "-" in a lineage III genome.

5.6 Phylogenomic Reconstruction

5.6.1 Neighbor-net Split Network

To reconstruct the phylogeny of *L. monocytogenes*, split networks were generated for the 26 *L. monocytogenes* strains based on the binary distributions (i.e. presence or absence) of 2,846 EGD-e protein-coding genes and 3,560 homologous groups (HGs). A maximum-likelihood method was used to compute gene content distance between all pairs of genomes [145], and the Neighbor-net algorithm [146] was used to generate the split networks (Figure 5.5). Split networks do not enforce a tree topology, and are therefore able to show incompatible phylogenetic signals due to horizontal gene transfers and/or recombination, evident as parallel edges in the network. Although Neighbor-net does not reveal the full recombination history, it generates a planar estimate that can be aesthetically displayed. The split networks generated by both EGD-e genes and homologous groups clearly separated the three major lineages of *L. monocytogenes*. However, the split network based only on EGD-e genes distorted the network topology and placed some LIII genomes in unreasonably long branches (Figure 5.5b), indicative of an inherent bias caused by a restricted set of loci used for phylogenetic reconstruction [147].

Of note in LI, the serotype 4b strain N1-017 contains considerably more strain-specific genes (i.e. 112 non-phage genes and 16 prophage genes) than any other strain analyzed in this study. N1-071 appears to be closely related to serotype 1/2b strains in the LI cluster, and is likely to represent an evolutionary intermediate between the split of serotype 4b and serotype 1/2b [65]. Of note in LII, four strains

F6900, F6854, J2818 and J0161 were previously traced back to a single food processing facility over a time span of 12 years [148]. F6900 (a human isolate) and F6854 (a food isolate) were associated with a case of sporadic listeriosis in 1988; J0161 (a human isolate) and J2818 (a food isolate) were implicated in a listeriosis outbreak in 2000. These four isolates are clustered closely on a single branch, indicative of a recent common ancestry.

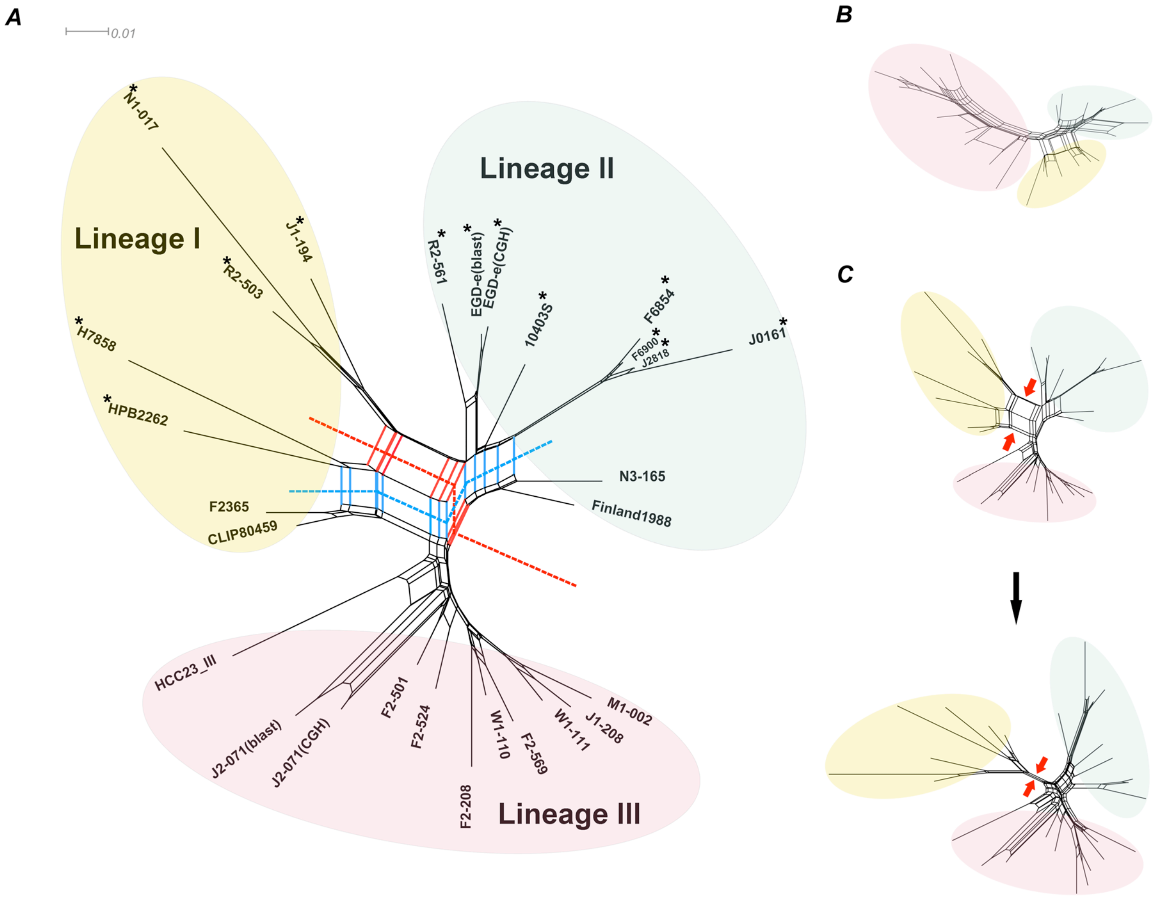


Figure 5.5: Split network of 26 *L. monocytogenes* genomes. a) Split network based on the presence or absence of 3,560 HGs in 7 LI, 9 LII and 10 LIII genomes. EGD-e and J2-071 were analyzed by both BLAST and CGH data. Two splits that caused the wide parallel edges at the root of LI and LII clusters are highlighted with blue and red dashed lines. Strains that carry the A118-like prophage in the *comK* gene are marked with *. b) Split network based on the presence or absence of 2,855 EGD-e core genes. c) Comparison of network topology before and after excluding the HGs largely responsible for the red and blue splits.

Although a Neighbor-net is a very limited estimate of the recombination history [146], it is an informative visual aid that highlights areas of the network most affected by incompatibilities. Parallel paths that reflect incompatible phylogenetic signals are most obvious at the base of branches leading to the split of LI and LII. These network-like structures are indicative of horizontal transfer (HGT) or parallel genetic loss—events that cause genetic convergence. The wide stem at the root of LI (indicated by arrows in Figure 5.5c) contains two primary splits. Each split grouped some LI strains with LII (dashed lines, Figure 5.5a), implying possible HGT or parallel gene loss between LI and LII. By sequentially excluding HGs with the largest differential abundance across each split, the splits could be removed while preserving the overall network topology (Figure 5.5c). This heuristic procedure narrowed the root of LI, mirroring the clonal structure of phylogenetic analysis based on house-keeping genes [149]. Genes responsible for the major splits are likely to be the source of the observed homoplasy (or convergent evolution). This possibility is further evidenced by an overrepresentation (38%) of prophage associated genes in the 124 HGs, compared to an average of 2% prophage genes in a typical *L. monocytogenes* genome.

To further explore the impact of lysogenic phages on the topology of the split network, we surveyed the distribution of all prophage regions in 26 *L. monocytogenes* genomes. A known A118-like prophage inserted in the *comK* gene [150] was common in five LI (R2-503, J1-194, N1-017, H7858 and HPB2262) and ten LII (EGD-e, R2-561, LO28, 10403S, F2-525, J2-003, F6900, F6854, J2818 and J0161) genomes. Insertion of this prophage caused the formation of a major split (indicated by blue dashed line) in Figure 5.5a, which separates strains with the *comK* prophage (indicated by a star sign) from those without it.

5.6.2 Neighbor-joining Tree

For a whole-genome analysis, traditional tree topologies can also accurately depict the consensus phylogenetic signal. Because all genes in the genome are considered, incompatibilities caused by horizontal transfer are drowned out by the much greater number of vertically inherited genes. Figure 5.6 shows a neighbor-joining (NJ) [151] tree for the 26 *L. monocytogenes* strains based again on the maximum-likelihood gene content distances. The NJ tree clearly delineates the three major lineages and the three subgroups in LIII. To test the compatibility between the CGH and tblastn datasets, both types of data for EGD-e and J2-071 were included in the analysis. The two data types for the same strain always clustered tightly together on a single branch, suggesting that the different data types are compatible for phylogenetic reconstruction.

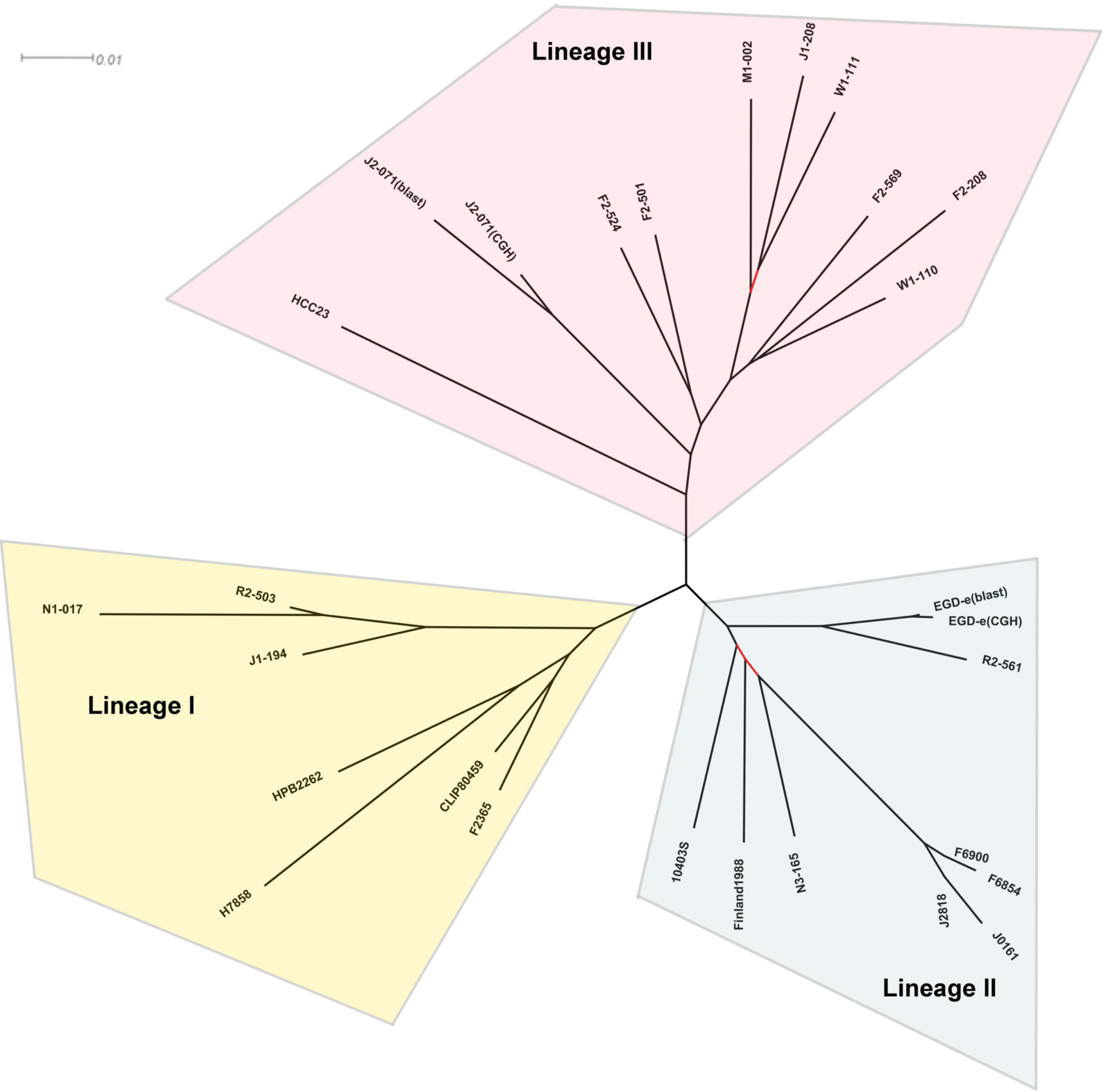


Figure 5.6: Neighbor-joining tree of 26 *L. monocytogenes* genomes. This tree was built based on the presence or absence of 3,560 HGs using the maximum-likelihood gene content method for 26 *L. monocytogenes* genomes. This tree was assessed by bootstrap analysis of 1,000 replicates. Braches with a bootstrap value lower than 70% are highlighted in red.

5.6.3 Elevated Diversity in Evolutionary Lineage III

Figure 5.7 shows a rooted NJ tree for the three LIII subgroups, using EGD-e as outgroup. HCC23 appears to be most closely related to LIIIA. Further evidence that links HCC23 to LIIIA is the rhamnose utilization gene cluster. This gene cluster is conserved in LIIIA and HCC23 but absent in LIIIB and LIIIC. The rooted NJ tree also suggests that LIII is paraphyletic and HCC23 possibly resembles an ancestral state of LIII. The emergence of 3 LIII subgroups is likely to be concomitant with stepwise genome reduction as observed in some non-pathogenic *Listeria* species, including *L. welshimeri* [134] and *L. seeligeri* [135].

A total of 206 genes, which are highly conserved in LI and LII, are phylogenetically informative for LIII (i.e. present or absent in at least one LIII strain). Figure 5.7c shows a heat map of the presence (red) or absence (black) of these genes in the ten LIII strains. Interestingly, sequential gene loss was observed in the order of LIIIA, LIIIC and LIIIB. Loss of LI and LII core genes was most significant in LIIIB. This LIII subgroup forms a deep branch in a split network (Figure 5.7b). It should be noted that the contribution of novel LIII genes to the phylogenetic reconstruction is likely to be underestimated due to the limited number of fully sequenced LIII genomes available at the time of this study.

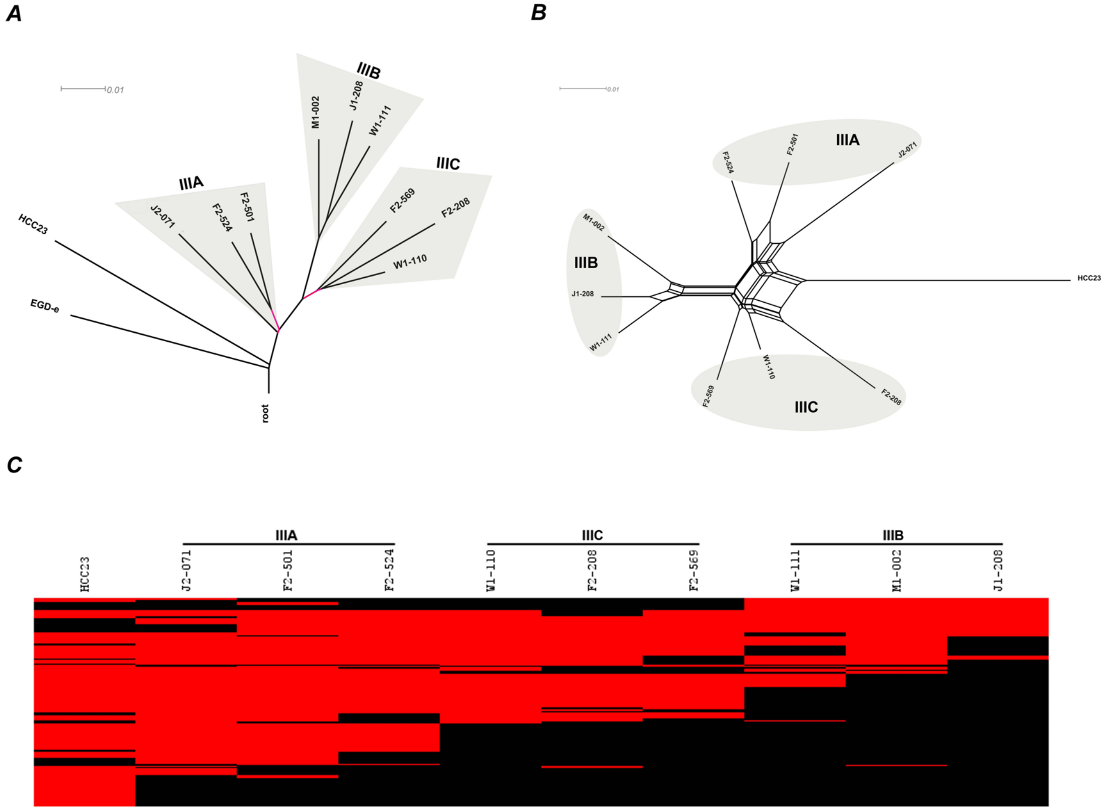


Figure 5.7: Phylogenetic analysis of the three LIII subgroups. a) A rooted tree shows the phylogenetic relatedness of the 9 LIII strains analyzed by CGH and 1 sequenced LIII strain HCC23. The tree was rooted by EGD-e and reconstructed based on the presence or absence of 3,560 HGs using the maximum-likelihood gene content method. Two branches with bootstrap values lower than 70% (1,000 replicates) are highlighted in red. b) Neighbor-net split network shows the phylogenetic relatedness of 10 LIII strains. c) A heat map shows the binary distribution of 206 LI and LII core genes in 10 LIII strains. Red color indicates a gene is present in a LIII genome; black color indicates a gene is absent.

To access the intra-lineage diversity from a gene content perspective, the accessory genes from F2365 (LI), EGD-e (LII), and J2-071 (LIII) were identified and their distributions in the three lineages were surveyed. Minimum spanning trees were used to visualize and compare the different distribution patterns of accessory genes in all three genetic lineages of *L. monocytogenes* (Figure 5.8). Accessory genes display similar distributions in most LI and LII strains. However, more complex and branched distributions were observed in LIII strains, clearly demonstrating an elevated genomic diversity in this rare *L. monocytogenes* lineage.

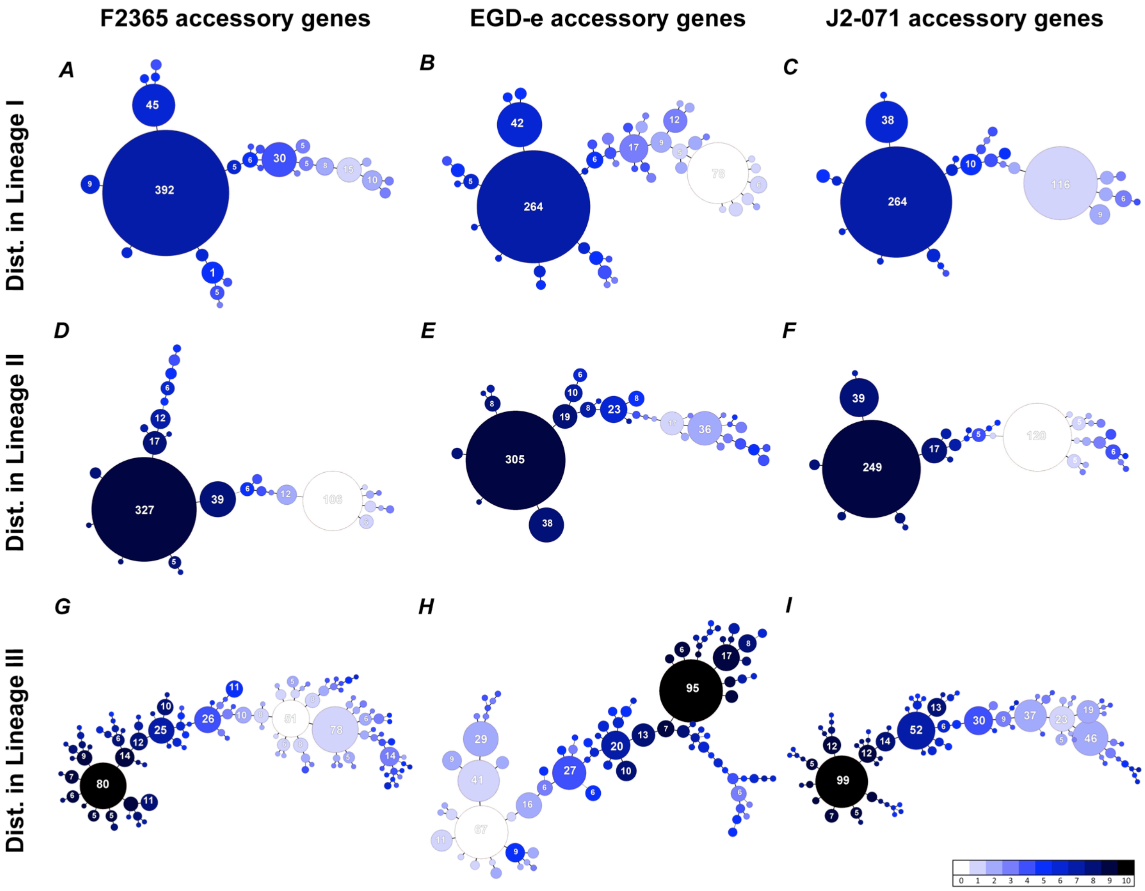


Figure 5.8: Accessory gene distribution in *Listeria monocytogenes* visualized with minimum spanning trees. A total of 576, 521 and 489 accessory genes were identified from F2365 (LI), EGD-e (LII), and J2-071 (LIII), respectively. The binary distribution of these accessory genes was surveyed in 28 *L. monocytogenes* genomes, including 4 newly sequenced strains. Each circle represents a group of accessory genes in F2365 (*A*, *D*, *G*), EGD-e (*B*, *E*, *H*), or J2-071 (*C*, *F*, *I*) that share a unique binary distribution (i.e. "1" for presence or "0" for absence) in all strains belonging to a specific lineage (i.e. I, II, or III). The size of each circle is proportional to the total number of genes that share the same binary distribution. Each circle is color-coded based on the number of *L. monocytogenes* strains (from 0 to 10, see color bar) that share the same distribution. This figure provides an overview of the genomic diversity of the three genetic lineages from a perspective of accessory gene presence or absence, in which LIII displays the most diversified gene content.

5.7 Analysis Methods

5.7.1 Bacterial Isolates and Hybridization

Table 5.1 lists the 26 *L. monocytogenes* strains analyzed in this study. As of November 2008, 20 sequenced *L. monocytogenes* strains were available and used for the pan-genomic array design. Array CGH was performed for 9 LIII strains representing 3 serotypes (4a, 4b, and 4c) and 3 subgroups (IIIA, IIIB, and IIIC). 4 additional isolates that were recently sequenced were incorporated in the pan-genomic and phylogenetic analysis. Bacterial strains were grown overnight in brain heart infusion (BHI) broth at 30°C. Genomic DNA was extracted and purified using MasterPure Gram positive DNA purification kit (EPICENTRE Biotechnologies, Madison, WI). Genomic DNA was labeled with Cy3 or Cy5 dyes prior to array hybridization.

Genomic DNA of each LIII strain was co-hybridized with that of EGD-e on a Roche NimbleGen 385K custom CGH array. Two dye-swap replicates were performed for each LIII strain/EGD-e pair to eliminate dye bias and test the array reproducibility. Genomic DNA labeling and array hybridization were performed at Roche NimbleGen, Inc. (Madison, WI). Hybridization results are available from GEO under accession number GSE20367.

5.7.2 Pan-genomic Array Analysis

To analyze this complex pan-genomic array, a new probe-based intensity classification scheme was devised that permits any locus to be classified based on the aggregated scores of its individual probes, without reference to control hybridization. Specifically, all raw signal intensities were first transformed to log values, then log intensities for replicate hybridizations were normalized using quantile normalization [152]. Replicates were combined at the probe level by taking the average of the normalized log intensities for each probe. Quantile normalization assumes similar intensity distributions, so to avoid cross-sample normalization bias, each strain was normalized and processed independently.

Because there was no one single reference to operate on, and to preserve sensitivity for small polymorphisms, intensity data was not smoothed or segmented. Instead, individual probes were each classified as present or absent using a minimum kernel density (MKD) method. MKD methods have performed well for the binary classification of both genes and segments [74,153], and here the idea is extended to the classification of individual probes. Because the array contains the full genetic diversity of *L. monocytogenes* and 4,300 random control probes, there is expected to be a significant fraction of both present and absent probe intensities for any *L. monocytogenes* sample. Therefore, the distribution of probe intensities is generally bimodal, and the minima between the present and absent peaks can be used as an effective threshold for binary classification. For each sample, the probability density function of the observed intensities was estimated using kernel density estimation and the central minima of this function identified as the optimal cutoff (Figure 5.9). This method was preferred because it is non-parametric, there is no potential normalization bias, it requires no training, and each sample can be processed independently without affecting the accuracy. It is also extremely flexible, in that a classification for any gene can be generated by aggregating the classifications of the probes targeting that gene. For this purpose, genes were scored by collecting all probes known to target a specific gene and computing the fraction of probes classified as present, the positive fraction (PF). A PF threshold of 0.6 was chosen by analysis of ROC curves for the EGD-e and J2-071 controls to minimize the total error rate (false-positive rate + false-negative rate) versus the tblastn 50% protein similarity threshold. PF was favored because it does not depend on cross-sample normalization, as would be necessary for an intensity threshold, and additional genomes can be analyzed independently without affecting accuracy. This makes it ideal for rapid and economical genotyping of novel isolates.

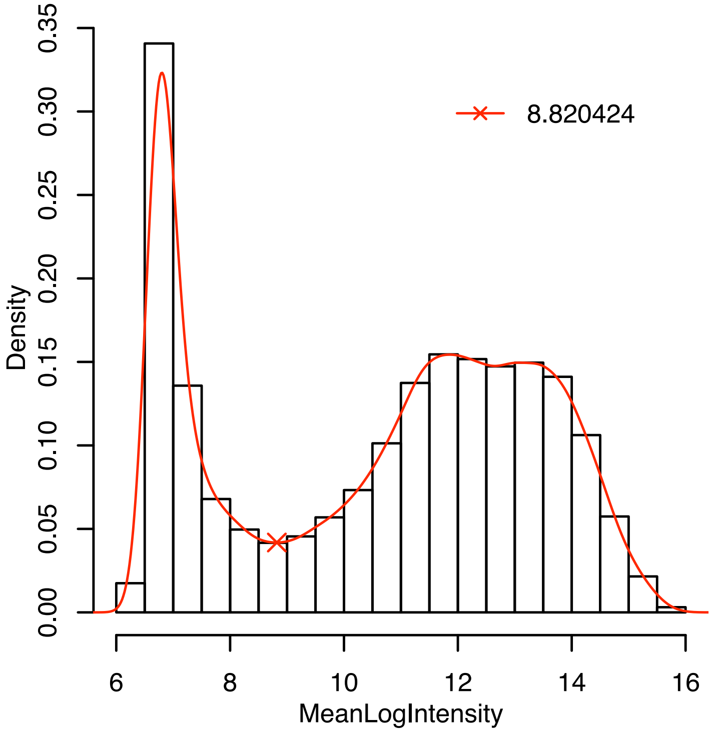


Figure 5.9: Probe intensity histogram overlaid with kernel density estimation (red) for sample J1-208, showing an optimal intensity cutoff of 8.82 at the minimum between the present and absent modes. Displayed distribution is for the mean intensities of the two quantile normalized replicates for strain J1-208.

5.7.3 Pan-genomic Sequence Analysis

Pan-genome analysis was performed using the methods introduced by Tettelin *et al.* [73], with modifications on the conservation threshold and permutation sampling. Annotated proteins for each genome were aligned to the six frame translations of all other genomes using tblastn. Query proteins were marked as present in a subject genome if the corresponding amino acid sequences aligned at ≥50% similarity with an E-value ≤10 -5, where “similarity” was defined as the number of positively scored residues divided by the length of the protein sequence. This threshold is more stringent than originally proposed in [73], but less stringent than those used in other studies (e.g. [120]). The 50% threshold was empirically selected as a compromise between tolerating draft genomes with fragmented annotations and avoiding false positive detections due to conserved domains and distant paralogs. A PF threshold of 0.6 was consequently chosen as an analogous threshold for the CGH results, as described above. Only genomes with fewer than 200 contigs were considered for the analysis.

The addition of an N_{th} genome was simulated by examining ordered combinations of N genomes. Due to the large number of available genomes, it was not feasible to consider all possible permutations as originally suggested. Instead, a randomly selected subset of 100,000 permutations was considered for the addition of each N, and the mean (or median) values were computed from this subset. For each permutation, the number of new genes found in the N^{th} genome G_N was computed as the number of proteins of G_N not present in any genomes G_i for $i=\{1,...,N-1\}$. The number of core genes was computed as the number of proteins of G_N present in all genomes G_i for $i=\{1,...,N\}$. Because gene sequences for the CGH strains are not known, EGD-e was set to be G_N for all permutations. The number of pan genes in a permutation of N genomes was computed by examining the genomes G_i in order from

1 to N. A gene in G_i was identified as a pan-gene if it was not present in any of the genomes G_j for all $j < i$.

The Gauss–Newton method implemented by the R function *nls* [87] was used to perform non-linear least squares regression on the mean and medians of the core genes, new genes, and pan genes distributions. According to [88], the number of new genes n expected to be discovered by sequencing an N^{th} genome was modeled by the power law function $n = \kappa N^{-\alpha}$, and the number of pan genes also by a power law $n = \kappa N^{\gamma}$. According to [73], the number of core genes was modeled by the exponential decay function $n = \kappa e^{-N/\tau} + \Omega$, where Ω describes the horizontal asymptote and therefore the core genes estimate. In all cases, the functions were fit to the mean or median values for all $N > 1$.

To accommodate false-negative errors introduced by sequencing gaps and weak hybridization signal, the originally proposed exponential decay function was modified with the addition of a fourth parameter to model the effect of a constant number of false-negatives with the addition of each genome, yielding:

$$n = \kappa e^{-N/\tau} + \Omega - N\beta$$

where the linear parameter β represents the number of core genes lost to false-negative errors for each N. Core gene loss due to false-negatives is not a truly linear phenomenon (e.g. sequencing gaps are not independent and the core genome can never be negative), but for a large core genome and a modest N it is a reasonable approximation that is easy to fit. To assure convergence of the optimization algorithm, β was first estimated via linear regression for $N \geq 15$, and this was used as the start estimate of β for the full model regression. The augmented model is useful in that the observed core genome size may be linearly decreasing (as is expected for draft genomes), but an estimate of the true core genome size Ω may still be recovered.

5.7.4 Homologous Group Identification

Homologous groups (HGs) were used for phylogenetic reconstruction and core genome estimation. HGs were identified by clustering a graph of protein similarity for all annotated protein-coding genes from the 18 high-quality *L. monocytogenes* genomes. A node was added to the graph for each one of the 52,776 annotated proteins. Edges were added between any two proteins with an alignment above the 50% similarity threshold. Unlike OrthoMCL, no orthology constraint was applied. Edges between any two similar proteins were added, including edges between proteins in the same genome. This was necessary due to the inability of CGH to accurately determine orthology. The MCL clustering algorithm was applied to this graph using an inflation parameter of 2.0. From this clustering, 3,744 HGs were identified, including strain-specific genes represented as singleton clusters. Some HGs, mostly singletons, were not represented on the array because additional genomes had been sequenced after the array design. A total of 3,560 HGs, represented on the array by at least one member gene, were used for the phylogenetic analysis.

For sequenced genomes, an HG was called present if at least one member protein of the HG aligned above the 50% similarity threshold. For CGH genomes, an HG was called present if at least one member gene of the HG hybridized with $PF \geq 0.6$. Results based on this threshold were converted to a unified binary table indicating gene presence or absence for all HGs in all genomes analyzed in this study. These binary vectors were used for measuring evolutionary distance using the maximum-likelihood measure of [145], and Neighbor-net split networks [146] and neighbor-joining trees [151] were built using the SplitsTree program [154]. Alternative parsimony methods failed to build reasonable trees, most likely due to the large number of incompatible splits caused by both horizontal gene transfer and errors in the data.

5.8 Discussion

5.8.1 Pan-genomic Comparative Genomic Hybridization

Pan-genome CGH was used in this study to compare *L. monocytogenes* genomes in pursuit of novel genes that potentially promote the fitness and virulence of LI and LII strains in human, as these strains are predominantly associated with human listeriosis. Phylogenomic concepts [155] guided the search for DDGs and to inferred the phylogeny for the species. Array CGH is ideal to serve the purpose of this study because it is cost-effective, accurate and highly reproducible. Compared to low-coverage sequencing, which often produces draft genomes with many gaps, CGH provides more rapid turnaround and more reliable inference of gene content. This was evident in the analysis, in that genomes with highly fragmented assemblies (e.g. F2-515 and J1-208) resulted in inaccurate core genome estimations and extremely long phylogenetic branch lengths, due to false-negative gene calls. The CGH approach circumvented such problems. Importantly, the pan-genome array design sampled the entire genetic repertoire of the species, thereby minimizing potential phylogenetic bias.

A particular challenge in this study was to unify the analysis of both genome sequence and CGH array data. The sensitivity of the two methods is fundamentally different. BLAST searches are capable of precisely measuring amino acid similarity and can identify orthologs and detect distant homologies. In contrast, DNA array hybridizations measure nucleotide conservation and are only capable of detecting highly conserved DNA sequences. In addition, hybridization gives no positional information and is non-specific, making it difficult to discriminate between paralogs. For this reason, homologous groups were used for gene content comparison, and permitted variant sequences to hybridize to their nearest neighbor in a group, rather than a single selected variant. Prior to implementing this method, there was tremendous detection bias in the CGH data. The HG method greatly increased the agreement between the array and BLAST detection strategies, which was critical for the phylogenetic analysis of the combined data.

5.8.2 Genes and sRNAs Associated with Niche-specific Fitness

The low frequency of LIII in human listeriosis can be partially explained by lack of or defective mutation in virulence factors. For instance, a novel streptolysin S-like hemolytic and cytotoxic virulence factor, listeriolysin S, was recently found to be exclusively present in LI strains [156]. This factor contributes to virulence of the pathogen in murine and human polymorphonuclear neutrophil-based assays [156]. Other studies also reported that premature stop codons are common in *inlA* in LIII strains [157,158,159,160]. Point mutations in *inlA* are presumably caused by localized recombination and lead to a truncated InlA protein and consequently a reduced invasion phenotype in human intestinal epithelial cells [157,158,159,160]. This pan-genome study uncovered 86 DDGs and 8 non-coding small RNAs that are absent or mutated in the largely uncharacterized LIII genomes. Most of these genes fall into the functional categories of cell wall structure, transcription regulation, and carbohydrate metabolism and transport. Such functions are likely to play critical roles in ecological fitness of *L. monocytogenes* in different environment and host niches. Genes involved in carbohydrate metabolism and transport stand out as the largest functional group of DDGs, implying that the capability of utilizing different carbon sources in the transmission and infection cycle contribute most to the predominance of LI and LII strains in human infections. In particular, PTS systems that are likely to confer niche-specific metabolic advantages are conserved in LI and LII but decayed or lost in LIII. For example, the fructose-like PTS components (*lmo2133–lmo2137*) are conserved in all LI and LII genomes but completely lost in LIIIB and LIIIC. This operon was postulated to have been acquired by *L. monocytogenes* through HGT from *Enterobacteriaceae* that cohabitate the GI tract of mammalian host [161]. A recent study of its homolog in extraintestinal pathogenic *E. coli* suggested that this operon promotes bacterial fitness against the stress in host serum and gut, and also enhances bacterial invasion in eukaryotic cells [162]—both are integral parts of listerial pathogenesis.

L. monocytogenes possess extraordinary capabilities for sustaining harsh conditions during its residency in the environment (e.g. it can utilize limited carbon source), in foods (e.g. it can resist salts and grow at refrigeration temperatures), and in parasitized hosts (e.g. it can escape from immune defense). During its passage through the human GI tract, *L. monocytogenes* is able to resist the antimicrobial effects imposed by gastric contents. Multiple genes involved in combating GI tract-related stresses, primarily gastric acid (*gadD1*, *gadT1* and the ADI system) and bile salts (*btlB* and *pva*), are missing in LIII. Lost of these genes may result in a defective phenotype in surviving the GI tract prior to invasive infection [136]. Also absent in most LIII genomes are a number of small regulatory RNAs (e.g. *rli29* and *rli48*) and transcription factors (e.g. *lmo2138* and *lmo2851*) that appear to be up-regulated in the human intestine [85]. It is reasonable to speculate that the human GI tract may act as a major barrier to prevent LIII strains from causing systematic infections. Epidemiological studies seem to support this speculation by collectively showing that gastroenteritis, rather than more severe listeriosis symptoms, is predominant among infected individuals [163,164,165]. Although intracellular strategies have been the primary focus in numerous studies of listerial pathogenesis, a few recent studies demonstrated that the GI passage has a fundamental impact on listerial pathogenicity [166,167]. Considering that most LIII strains possess virulence factors related to its intracellular lifestyle and are cytopathogenic [109], the inability to survive in the GI tract becomes a more reasonable explanation for the overall rarity of LIII in human listeriosis.

5.8.3 Core and Pan-genomes of *Listeria monocytogenes*

L. monocytogenes core-genome consists of approximately 2,330 to 2,456 genes and the pan-genome encompasses over 4,052 genes. Compared to several other bacterial species, *L. monocytogenes* has relatively higher proportions (about 80%) of core genes shared by individual genomes (Table 5.6), which in turn reflects lower intraspecies genomic variability. This is consistent with the low rates of recombination in this bacterial species [157]. Despite the perceived high genomic synteny, *L. monocytogenes* possesses considerably diverse pan gene reservoir and displays biased distribution of accessory genes across major evolutionary lineages.

Table 5.6: Summary of other pan-genomic studies.

Species	No. Genomes[1]	Pan genome[2]	No. core genes	No. pan genes	Avg. no. genes	% Core genes	Cutoff[3]	Ref
Escherichia coli & Shigella	20	Open	1976	>17838	4700	42%	80/80	[121]
Escherichia coli	17	Open	2200	>13000	5020	44%	0.8 BSR	[120]
Escherichia coli	32	Open	1563	>9433	4537	34%	50/50	[74]
Haemophilus influenzae	13	Finite	1461	4425–6052	1970	74%	70/70	[118]
Listeria monocytogenes	26	Open	2350–2450	>4000	2978	80%	0.5 SSR	-
Neisseria meningitidis	7	Open	1333	>3290	1963	68%	50/50	[119]
Streptococcus agalactiae	8	Open	1806	>2750	2245	80%	50/50	[73]
Streptococcus agalactiae	8	*Open	1472	*>2800	2198	67%	1e-5 E-value	[116]
Streptococcus pneumoniae	17	Finite	1380	5100	2438	57%	70/70	[117]
Streptococcus pyrogenes	11	*Closed	1376	*2500	1878	73%	1e-5 E-value	[116]

All numbers are estimates in this table.

[1] Only studies including more than five strains are shown.

[2] Pan-genome growth behaviors as described by the authors. *Estimated from figures, but not explicitly stated.

[3] Cutoff values and methods for defining core and pan genes vary widely across the different studies. This column only gives a rough summary of the similarity cutoff. Cutoffs of the form *I*/*L* indicate a minimum BLAST hit of *I*% similarity over *L*% of the protein length. BSR is Blast Score Ratio [32]. SSR is the similarity score ratio used in this study, similar to BSR.

5.8.4 Implications for Niche Adaptation

The extensive and high-resolution coverage afforded by the pan-genome approach allowed both robust phylogenetic reconstruction and systematic examination of specific genomic features associated with individual strains and lineages. Some incompatible phylogenetic signals confounding the genealogy of LI and LII strains were traced back to prophage genes. The *comK* prophage regions in different *L. monocytogenes* genomes display significant sequence variations (Figure 5.10). Such variations may be a result of prophage decay, recombination that have accumulated in the remnants of common prophage ancestor(s), or multiple lysogenization of different bacteriophages at the same genomic location. Temperate listerial phages generally have strict host specificity as defined by bacterial surface receptors, and do not cross infect between two serogroups (serotypes 1/2 and 3 in one group; serotypes 4, 5 and 6 in the other). For example, *Listeria* phage A118 has been shown to be specific to serotype 1/2 or LII [168,169]. The presence of A118-like prophage in serotype 4b or LI strains (N1-017, H7858, and HPB2262) can be reasonably interpreted by the integration of A118-like phages into the ancestral strain(s) prior to the divergence of contemporary lineages. *L. innocua* also harbors a *comK* prophage, suggesting that the phage integration may even have preceded the speciation split between *L. monocytogenes* and *L. innocua*. Phages have been well recognized as the major contributors of important biological properties (e.g. virulence factors) in many bacterial species [170,171]. The functional impact of bacteriophages on the biology of *L. monocytogenes*, if any, has yet to be determined.

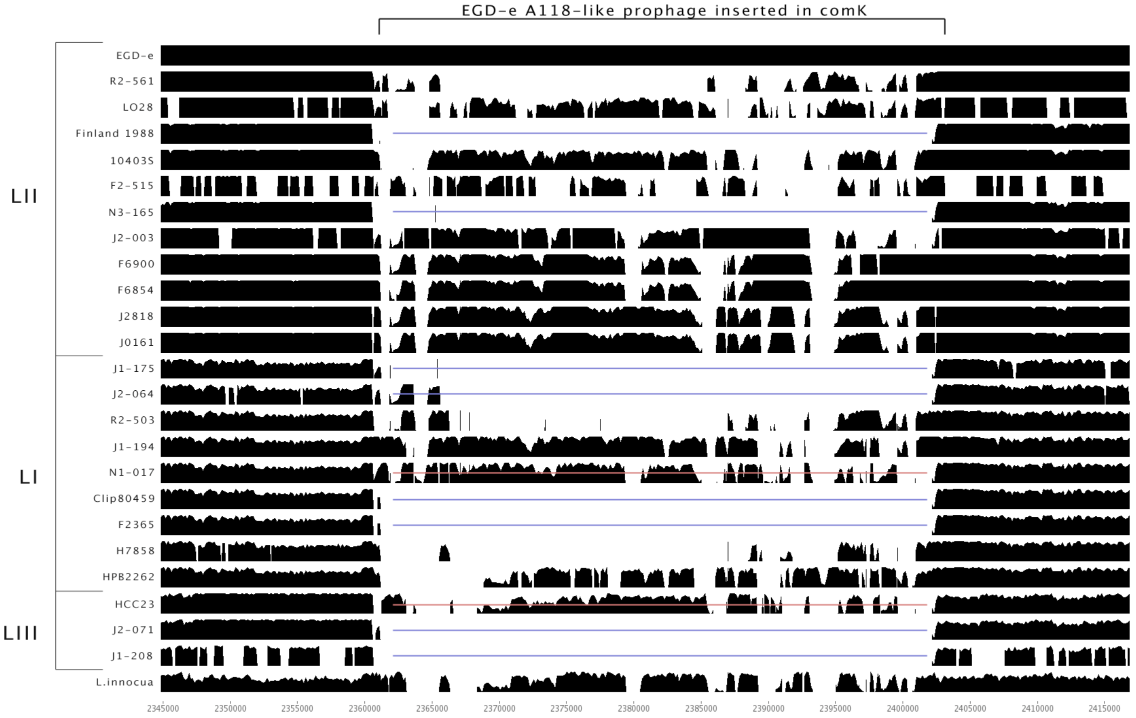


Figure 5.10: Alignment of A118-like prophage in different *L. monocytogenes* lineages. The x-axis gives the location on the EGD-e chromosome, and for each strain, windowed alignment identity is given on a scale of 50–100% identity on the y-axis. Strains that show no homology to the EGD-e A118-like prophage are struck through in blue. Strains which do show homology to the prophage, but the prophage is inserted somewhere other than *comK*, are struck through in red (N1-017, HCC23). This plot illustrates some interesting phylogenetic incompatibilities. For example, based on whole-genome analysis, the nearest phylogenetic neighbor to EGD-e is R2-561. Yet the *comK* prophage in nearly all other strains appears more similar to EGD-e than does the prophage in R2-561, which has identity <50% for most of its length.

5.9 Summary

Intraspecific variations in host preference, ecological fitness and virulence potential are common in many pathogenic species. This study used a pan-genomic approach that combines *in silico* comparative genomic analysis and high-density CGH arrays to explore the genomic diversity of *L. monocytogenes*. Based on our results, one *L. monocytogenes* strain carries about 75% of the pan genes of this species. That said, experiments based on a single reference strain may not adequately sample the total genetic repertoire and not fully interpret the versatile biology of *L. monocytogenes*. A defined species core genome may supplement a new genomic criterion for taxonomic classification of *L. monocytogenes*, as some traditional methods are often inconclusive and controversial. Genes and regulatory RNAs identified from this study may help elucidate the predominant association of *L. monocytogenes* LI and LII with human listeriosis. The pan-genomic approach described here can also be used to explore the genomic diversity in other pathogenic species, as such information would help better understand the intraspecific variations in virulence, and the ecology, epidemiology and evolution of microbial pathogens.

5.10 Author Contributions

A version of this chapter has been submitted for publication:

Deng X*, Phillippy AM*, Li Z, Salzberg SL, Zhang W. Probing the pan-genome of *Listeria monocytogenes*: implications for intraspecific niche expansion and genomic diversification. *In preparation*. **Equal contribution*.

This text was coauthored, with equal contribution, by Xiangyu Deng. I designed and implemented all computational methods; designed the array; performed the array, phylogenomic, core and pan-genome, and sequence analysis; and drafted the corresponding sections of the text. Xiangyu Deng conceived, performed, and interpreted the biological experiments; performed the DDG and lineage-specific gene analysis; and drafted the corresponding sections of the text and the introduction. Zengxin Li contributed to the DDG and lineage-specific gene analysis. Wei Zhang conceived the experiments, coordinated the project, and edited the text.

Chapter 6

Whole-genome Assembly Validation[†]

When PCR and microarray based methods do not provide suitable accuracy or detail, whole-genome sequencing is the comprehensive diagnostic capable of decoding every nucleotide of a genome. However, it has been increasingly common to sequence genomes only to draft quality, bypassing the critical stages of manual validation and genome closure. For diagnostic and forensic purposes, the accuracy of these sequences is critical.

This chapter presents a collection of tools aimed at automated genome assembly validation, formalizes several mechanisms for detecting mis-assemblies, and describes their implementation in an automated validation pipeline. The application of this pipeline is demonstrated in both bacterial and eukaryotic genome assemblies, which highlights several assembly errors in both draft and finished genomes. The software described is compatible with common assembly formats and is released open-source (http://amos.sourceforge.net).

[†] This chapter includes previously published work with multiple authors. See Section 6.6 for details.

6.1 Background

Sequence assembly errors exist in both draft and finished genomes. Since the initial “draft” sequence of the human genome was released in 2001 [172,173], great effort has been spent validating and finishing the official sequence. During this process, it became clear that the original draft sequences were not entirely accurate reconstructions of the genome [174,175,176,177]. It was also reported in 2004 that “finished” human bacterial artificial chromosome (BAC) sequences contained a single base-pair error per every 73 Kbp of sequence and more significant mis-assemblies every 2.6 Mbp [174]. Some errors had left large stretches of sequence omitted, rearranged, or otherwise deformed. After five more years, the human genome was nearly complete, however the validation and finishing was a largely manual, and expensive, process requiring additional laboratory work and sequencing.

For many other genomes, cost prohibits manual sequence validation, and the genomes are often left as draft assemblies. Such sequences likely contain many errors, and recent calls for caution have been made regarding assembly quality [178]. Too often, assembly quality is judged only by contig size, with larger contigs being preferred. However, large contigs can be the result of haphazard assembly and are not a good measure of quality. It has been difficult to gauge assembly quality by other means, because no automated validation tools exist. The following sections catalog the identifiable anomalies that result from incorrect reconstructions of a genome, and describe a software pipeline for validating the output of assembly programs.

6.1.1 Double-barreled Shotgun Assembly

Shotgun sequencing, the most widely used DNA sequencing technique to date, involves three major steps: (i) the DNA is randomly sheared into fragments (shotgun step); (ii) the ends of each fragment are sequenced, resulting in two reads per fragment (double-barreled sequencing step); and (iii) the original DNA sequence is reconstructed from the reads (assembly step). Newly emerging sequencing technologies also follow this general model, albeit with different strategies for each step. The first two steps are highly automated, however the assembly step remains a difficult challenge for any sequencing technology. Assembly would be a trivial process if each read had a unique placement, however all but the simplest organisms contain duplicated sequences (repeats) throughout their genome. These repeats confuse the assembly process, since reads originating from distinct copies of the repeat appear identical to the assembler. Additionally, for near-identical repeats, it is difficult to differentiate sequencing error from the polymorphism between repeat copies. This may cause an assembler to incorrectly place repetitive reads, resulting in mis-assembly. The pairing of reads sequenced from opposite ends of a same DNA fragment (mate-pairs, or paired ends) helps to disambiguate read placements within and around repeats, as show in Figure 6.1a where ambiguous placements can be resolved by reads whose mates are anchored in unique sequence.

In a correct assembly, the layout of the reads, and implicitly, the layout of the original DNA fragments, must be consistent with the characteristics of the shotgun sequencing process used to generate the data. In general, a correct assembly must satisfy the following constraints:

1. The sequences of overlapping read must agree. Exceptions: sequencing errors, polyploid organisms, and the assembly of mixed samples such as non-clonal or out-bred organisms.
2. The distance between mated reads must be consistent with the size of the fragments generated from the random shearing process. Exceptions: chimeric DNA fragments.
3. Mated reads must be oriented towards each other, i.e. they must come from opposite strands of the sequenced DNA. Exceptions: chimeric DNA fragments, and alternative pairing methods (e.g. transposon libraries).
4. The placement of reads throughout the assembly must be consistent with a random shearing process, represented mathematically as a Poisson process [179]. Exceptions: cloning or sequencing biases.
5. All reads provided to the assembler must be consistent with the resulting assembly, i.e. every read must perfectly match at least one location in the reconstructed genome. Exceptions: sequencing errors, incomplete trimming of the sequencing vector, and the presence of contaminants.

All five of these constraints are subject to some degree of inaccuracy, as evidenced by the exceptions indicated above. A single violation is, therefore, not usually conclusive of mis-assembly. Instead, multiple, coinciding constraint violations need to be observed in order to infer the presence of an error in assembly. The following section describes the primary types of mis-assemblies and the pattern of constraint violations they exhibit.

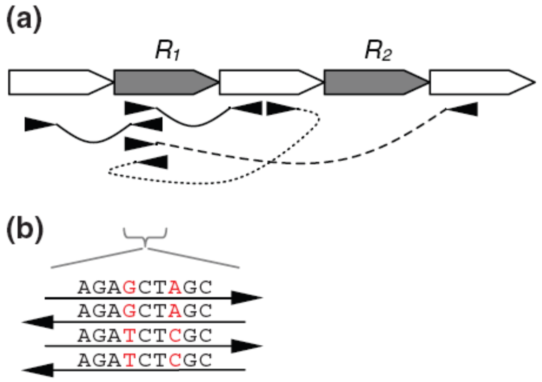


Figure 6.1: Mis-assembled sequencing reads caused by the two copy repeat *R* and leading to (a) unsatisfied mate-pairs and (b) correlated SNPs. Unique sequence is shown in white and repetitive sequence in gray. Example mate-pairs are drawn as connected arrowheads. Properly oriented mates point towards each other, and properly sized pairs are connected with a solid line. All mates can be satisfied and the correlated SNP removed if the bottom two reads in R_1 are moved to R_2.

6.1.2 Mis-assembly Signatures

The majority of mis-assemblies fall into two generalized categories: (i) repeat collapse and expansion; and (ii) sequence rearrangement and inversion. Each type has distinct mechanisms for mis-assembly and result in different signatures. The first type of mis-assembly results from incorrectly gauging the number of repeat copies in a genome and including too few or too many copies. Differences in copy numbers of certain repeats are known to cause phenotypic differences between organisms (e.g. Huntington's disease [180]), therefore a correct assembly of such regions is essential. The second type of mis-assembly results from shuffling the order of multiple repeat copies, thereby rearranging the unique sequence in between. This type of mis-assembly, if uncaught, could be misinterpreted as a biological rearrangement event. There is a chance such false conclusions have already been drawn due to mis-assembled genomes, and therefore the mechanisms and signatures of these mis-assemblies need to be examined in more detail.

In both collapse and rearrangement events, reads may be placed in the wrong copy of a repeat. Small differences between repeat copies, often single nucleotide polymorphisms (SNPs) caused by mutations that arose in the different copies independently, are useful indicators of collapsed or otherwise mis-assembled repeats. While disagreements due to sequencing errors tend to occur at random, the differences caused by mis-assemblies can be identified by their correlated location across multiple reads (Figure 6.1). Some correlated SNPs may also occur due to heterogeneous sequencing samples or sequence-specific lab errors, and therefore correlated SNPs by themselves are not always sufficient evidence of mis-assembly.

6.1.2.1 Repeat Collapse and Expansion

In the case of a repeat collapse, the assembler incorrectly joins reads originating from distinct repeat copies into a single unit (Figure 6.2). The opposite occurs in an expansion, where extra copies of a repeat are included in the assembly. These often result in a greater (or lesser) density of reads than is expected from the random shotgun process. A missing repeat copy causes reads to “pile up” in the remaining copies, thereby increasing read density. For example, in a genome sampled at 8-fold coverage with reads of 800 bp in length, the reads are expected to be placed at approximately 100 bp increments throughout the genome. The collapse of a two copy repeat results in an even denser packing of the reads in the single remaining copy—within the collapsed repeat the reads are spaced by roughly 50 bp and the depth of coverage (number of reads spanning a specific location) is increased to about

16-fold. The reverse is true for an expansion mis-assembly, where the read density drops below normal coverage.

In the case where two repeat copies are adjacent to each other, i.e. a tandem repeat, the reads that span the boundary between the two copies cannot be placed in the collapsed assembly. These reads only partially align to the assembly and exhibit an identifiable mis-assembly signature where they appear to wrap-around the boundary of the repeat. In addition, mate-pairs spanning the boundary between the two copies, but internal to the tandem, also appear to wrap around and mates spanning the tandem are shorter than expected (Figure 6.2b). For expansions, spanning mates appear stretched. When two repeat copies are separated by a unique region, a collapse forces the intervening section of DNA out of the assembly, leading to the creation of two separate contigs. Any mate-pairs that were spanning one of the repeat copies now link from the excised contig to the middle of the collapsed contig (Figure 6.2d). An insertion results in a similar signature, with mates spanning the insertion boundary linking to separate contigs. In general, any non-overlapping placement of two contigs with respect to each other results in the violation of mate-pair constraints, indicating the presence of a mis-assembly.

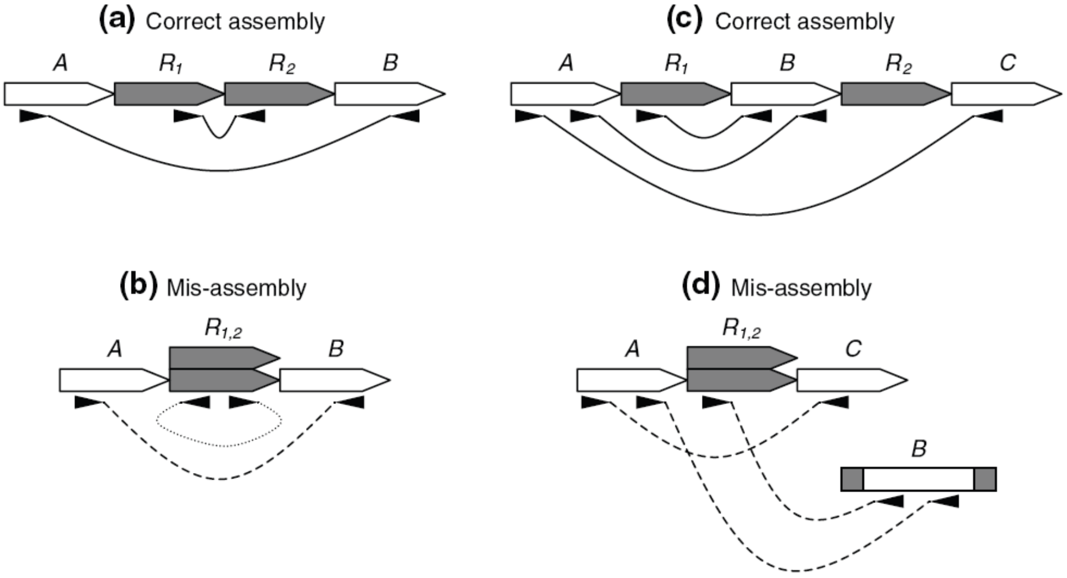


Figure 6.2: Mate-pair signatures for collapse mis-assemblies. (a) Two copy tandem repeat *R* shown with properly sized and oriented mate-pairs. (b) Collapsed tandem repeat shown with compressed and mis-oriented mate-pairs. (c) Two copy repeat *R*, bounding unique sequence *B*, shown with properly sized and oriented mate-pairs. (d) Collapsed repeat shown with compressed and mis-linked mate-pairs.

6.1.2.2 Rearrangements and Inversions

Even when an assembler correctly gauges the number of repeat copies, thereby avoiding the situations described above, mis-assemblies are still possible. Such a situation is shown in Figure 6.3, where, by incorrectly redistributing reads between the three copies of repeat *R*, the regions *B* and *C* of the genome have been swapped. Inversions are a special case of rearrangement, occurring when two repeat copies are oriented in opposite directions, thereby allowing the intervening region to be inverted (Figure 6.4). These "inverted" repeats can easily confuse the assembler, and can also result in genomic rearrangements in vivo, such as those detected within the plasmids of *Bacillus anthracis Ames* [4]. In the case of mis-assembly, heterogeneities may result within the mis-assembled repeat copies, due to mis-placed reads, unless the repeat copies are identical. In addition, mate-pair constraints are violated for any mate-pairs spanning the repeat unit. If the repeat is not spanned by mate-pairs, this class of mis-assembly is harder to detect, and it is sometimes possible to mis-assemble the genome without violating a single mate-pair constraint. While a random placement of the reads among repeat copies would result in violations, assembly programs often place the reads such that the constraints are satisfied, thereby obscuring the mis-assembly.

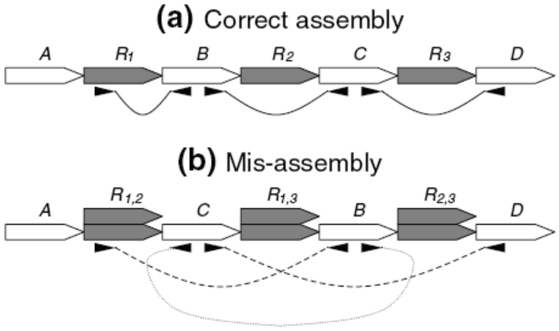


Figure 6.3: Mate-pair signatures for rearrangement mis-assemblies. (a) Three copy repeat *R*, with interspersed unique sequences *B* and *C*, shown with properly sized and oriented mates. (b) Mis-assembled repeat shown with mis-oriented and expanded mate-pairs. The mis-assembly is caused by co-assembled reads from different repeat copies, illustrated by the stacked repeat blocks.

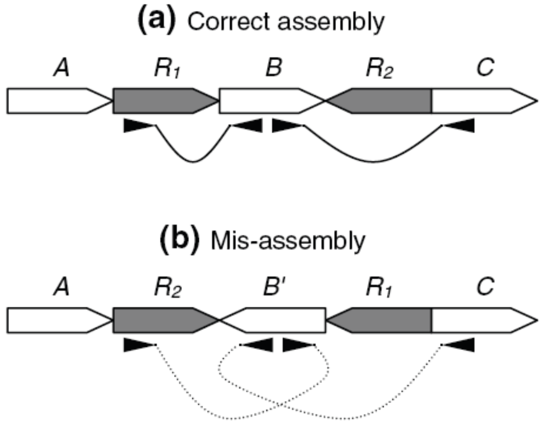


Figure 6.4: Mate-pair signatures for inversion mis-assemblies. (a) Two copy, inverted repeat *R*, bounding unique sequence *B*, shown with properly sized and oriented mate-pairs. (b) Mis-assembled repeat shown with mis-oriented mate-pairs.

6.1.3 Prior Work

Gene Myers' original formulation of the assembly problem stated that an assembly of a genome must match (in terms of the Kolmogorov-Smirnoff test statistic) the statistical characteristics of the process used to generate the data [181]. This is the first formulation of the assembly problem that explicitly takes into account the presence of repeats in genomes. Furthermore, this formulation provides a theoretical framework for developing assembly validation tools. A simple version of this approach, the arrival-rate statistic (A-statistic), is used within Celera Assembler to identify collapsed repeats [182].

The validation of genome assemblies was originally done manually, in conjunction with genome finishing efforts aimed at generating the complete sequence of organisms. Validation software was generally provided as an add-on to assembly editors like Consed [183], Staden package [184], or TIGR Editor (in-house software used at The Institute for Genomic Research). New interest in developing tools for assessing the quality of assemblies was spurred by the race to finish the human genome, in particular by the competition between the publicly led effort [172] and the private challenger Celera Genomics [173]. The ensuing controversy and flurry of papers comparing the two assemblies underscored the absence of objective and reliable tools for assembly validation. Eventually, the human assemblies were verified through comparisons to a collection of independently generated data such as finished BAC clones [185], gene content [186,187], and (at a lower resolution) genomic physical maps [172,173,188].

Such comparative validation methods have limited applicability. First, they rely on the availability of a "gold standard" provided by independently generated and often manually curated data. Second, these methods can only detect mis-assemblies covered by the sparse curated data. A more general approach utilizes just the assembly data itself, such as the constraints imposed by the mate-pairs, whose placement within the assembly must be consistent with the characteristics of the shotgun process. For example, a visual display of mate-pairs, the clone-middle-plot, was used to compare the two different assemblies of the human genome [189], and the popular assembly viewer/editor Consed [183] includes the means to explore the placement of paired reads along the genome as a tool for identifying mis-assemblies. An assembly viewer I co-developed, Hawkeye [190], presents the assembly as a tiling of paired reads, and provides several visualization options aimed at highlighting possible assembly problems. An integrated analysis of mate-pairs is built into the quality control module of the Arachne assembler [191,192]. The Arachne approach detects clusters of unsatisfied mate-pairs and low quality bases to estimate the probability of mis-assembly for each region of the assembly. In addition, two standalone programs are available for mate-pair based evaluations: BACCardI [193] allows the user to visualize the placement of mate-pairs along the genome and highlights those mate-pairs that are incorrectly placed with respect to each other, and TAMPA [194] uses a computational geometry algorithm to identify clusters of mis-mated reads that are characteristic of a mis-assembly.

Despite its many benefits, mate-pair based validation may produce many false positives due to the inherent inaccuracy in the experimental protocols. For example, in a correct assembly many mate-pairs would be characterized as incorrect, specifically those representing the tails of the mate-pair size distribution. This problem can be alleviated using statistical hypothesis testing, an approach used by the compression-expansion (CE) statistic [195]. In short, for every position in the genome, the CE statistic represents the deviation—in number of standard errors—of the observed mean mate-pair size from the mean size of the shotgun library (the statistical Z-test). A CE value near zero indicates the local distribution of sizes is in agreement with the global distribution, while large (e.g. greater than three) negative (positive) values indicate the presence of a compression (expansion) in the assembly. This statistic is less sensitive to the variance of mate-pair sizes, and therefore much more sensitive in identifying true errors.

An alternative approach to mis-assembly detection and resolution is taken by DNPTrapper [196]. This tool focuses on the heterogeneities between co-assembled reads to detect collapsed repeats, and provides an interface for manually separating the individual copies, using the Defined Nucleotide Position framework of Tammi *et al.* [197]. Another sequence based approach introduced by Kim *et al.* examines the distribution of sequences within all reads to identify repetitive, and therefore difficult to assemble, regions [198].

Despite their utility, none of the tools described above take into account more than one measure of assembly correctness. The next section describes *amosvalidate*, the first integrated pipeline for assembly validation that combines multiple observations and validation techniques to more accurately detect mis-assemblies. This comprehensive approach increases the sensitivity and specificity of mis-assembly detection, and focuses validation on the most probable mis-assemblies. Regions identified as mis-assembled are output in AMOS message format, thereby enabling the integration with other validation pipelines, as well as manual inspection with the Hawkeye assembly visualization tool.

6.2 Assembly Validation Methods

Violations of the five basic rules described in the Introduction are most commonly caused not by mis-assemblies, but by statistical variation or errors in the underlying data provided to the assembler. The high-throughput biochemical processes used to sequence genomes are error-prone, leading to non-random coverage across the genome, sequencing errors, and mis-paired reads. Furthermore, experimental measurements (e.g. mate-pair sizes) are inherently noisy. Separating such experimental artifacts from errors introduced by mis-assemblies is one of the main requirements of a robust validation pipeline. To reduce the effect of these errors on the analysis, multiple sources of evidence must be combined to increase the specificity of mis-assembly detection. In addition, certain types of mis-assembly can only be detected by specific methods, while the sequencing strategy employed may restrict the types of information that can be used for validation (e.g. many emerging sequencing technologies do not yet generate mate-pair information). The remainder of this section will describe validation techniques based on several measures of assembly consistency and how these measures can be integrated to reveal assembly errors.

6.2.1 Mate-pairs

The mate-pair validation component of the pipeline separately identifies the four types of mis-mated reads: (i) mates too close to each other; (ii) mates too far from each other; (iii) mates with the same orientation; and (iv) mates pointing away from each other. Reads with mates not present in the assembly or whose mates are present in a different contig are also reported. In order to reduce the impact of noise in the underlying data, multiple mate-pair violations must co-occur at a specific location in the assembly before reporting the presence of an error. In addition, the CE statistic described in the Introduction aids in the identification of clusters of compressed or expanded mate-pairs.

The actual size of shotgun libraries is sometimes mis-estimated by sequencing centers; therefore, a mechanism to re-estimate the library parameters on the basis of mate-pairs that are co-assembled within a contig is required. Reads that occur too close to the end of a contig may bias the distribution in favor of short mate-pairs (the mate-pairs at the upper end of the distribution would fall beyond the end of the contig and therefore not contribute to the calculations) and are therefore ignored. Specifically, reads are ignored if they are closer than $\mu + 3\sigma$ from the end of the contig when re-estimating the parameters of a library with mean μ and standard deviation σ. It is often necessary to iterate this process a few times until convergence. The size of a library is re-estimated only if the size of a sufficient number of mate-pairs can be estimated and only if either the mean or the standard deviation change significantly from the original estimate.

In addition to mate-pair violations, regions of inadequate depth of coverage are identified, as well as regions that are not spanned by any valid mate-pair (i.e. zero fragment coverage). The latter may represent situations where non-adjacent regions of the genome were co-assembled across a repeat. When computing fragment coverage the reads sequenced from each fragment are excluded from consideration. This is necessary in order to make the distinction between read and fragment coverage at a specific location. By this definition, the read coverage cannot drop below one within a contig, but the fragment coverage can be as low as zero, indicating the absence of long-range support for this region of the contig. At the typical depths of read coverage used in sequencing, each location in the genome is generally well covered by mate-pairs.

6.2.2 Repeat Statistics

Most mis-assemblies are caused by repeats, therefore, understanding the repeat structure of a genome can aid in the validation of its assembly. Some repeats can be found by aligning the assembled contigs against each other and identifying duplicated regions. Tools like Vmatch [26] and Tandem Repeat Finder [199] can be used for the *de novo* identification of repetitive regions in the assembly, which can then be examined for correctness. This approach, however, is not appropriate for all types of mis-assemblies. For example, the complete collapse of a two copy tandem repeat into a single copy cannot be detected by comparative means.

For validation purposes it is not sufficient to simply locate the repeats, rather it is more essential to identify those repeats that have been assembled incorrectly, especially those repeats that cannot be identified through comparative analysis.

Specifically, amosvalidate identifies regions of the genome that are over-represented in the set of reads, yet appear unique when examining the consensus sequence generated by the assembler. This is achieved by comparing the frequencies of *k*-mers (*k*-length words) computed within the set of reads (K_R) with those computed solely on the basis of the consensus sequence (K_C). K_R is the frequency of all *k*-mers inside the clear range of all reads; and K_C is the frequency of all *k*-mers across the consensus sequence of the assembled contigs. The forward and reverse complement of each *k*-mer are combined into a single frequency. The normalized *k*-mer frequency $K^* = K_R / K_C$ is computed for each *k*-mer in the consensus, where a deviation from the expected K^* (in a correctly assembled region, K^* should approximately equal the average depth of coverage) reveals those repeats likely to be mis-assembled. For example, K_R measured across a two-copy repeat is $2c$ regardless of whether the assembly is correct or not. If the repeat is correctly assembled into 2 distinct copies $K_C = 2$ and therefore $K^* = c$. If instead the repeat is collapsed, then $K_C = 1$ and $K^* = 2c$ indicating the presence of a mis-assembly. This approach is particularly powerful when used in conjunction with the technique described below for identifying dense clusters of SNPs because the two methods are complementary. SNP based detection will find collapsed, heterogeneous repeats, while K^* will reveal collapsed, identical repeats.

6.2.3 Coverage Statistics

As described in the introduction, the collapse of a repeat results in an increase in the depth of coverage. This characteristic signature can therefore be used to detect the presence of mis-assemblies. For short repeats with low copy number (e.g. 2-copy repeats), this effect cannot be distinguished from the variation in coverage caused by the randomness of the shotgun sequencing process, limiting the applicability of this method to repeats that occur in many copies throughout the genome, or to relatively long stretches of repetitive DNA (sustained deviations from the average depth of coverage are unlikely to occur by chance). The significance of observing a certain level of over-representation, given the parameters of the shotgun process, can be calculated through statistical means (see the A-statistic used by Celera Assembler [182]).

6.2.4 Identifying Micro-heterogeneities

Under the assumption of a random distribution of sequencing errors, and an independent random sampling of the genome during the shotgun process, it is unlikely that any two overlapping reads have sequencing errors at the same consensus position. While there are several examples of sequence-dependent sequencing errors that invalidate the assumption of independence between errors occurring in different reads, these assumptions are true for the vast majority of sequencing errors. Also, the following discussion assumes the genome being sequenced represents a single clonal organism. The assembly of non-clonal bacterial populations or heterozygous eukaryotes is characterized by frequent heterogeneities between co-assembled reads. Such situations are often known *a priori* and the validation pipeline can be adjusted accordingly.

As described in the introduction, mis-assemblies often result in the presence of micro-heterogeneities (SNPs) that are correlated across multiple overlapping reads. Identifying such polymorphisms can, therefore, indicate potential errors in the assembly. To identify mis-assembly induced SNPs, and distinguish them from simple sequencing errors, this technique leverages the base quality values provided by the sequencing software. The phred quality values [200] for example, represent the log-probability of error at every base in the sequence. Under the assumption of independence of errors across reads, these values may be summed to estimate the probability of observing multiple correlated errors at a specific location in the assembly, and mark as polymorphism those locations where this probability exceeds a specific threshold. For example, the probability of error for two reads reporting the same base, each with a quality value of 20, is equivalent to the probability of error for a single base with a quality value of 40 [P(error) = 1/10,000]. This is, in essence, the same approach used by genome assembly software in assigning quality values for the consensus sequence [201]. For each heterogeneous column of the multi-alignment, reads are grouped into "alleles" by which nucleotide they report. The quality values for each read in an allele are summed, and if two or more alleles have a quality value of 40 or greater (by default) the difference is marked as a SNP. For a concrete example, if two reads report a 'C' each with quality 25, and three reads report a 'G' each with quality 20, the qualities of the alleles are 50 and 60 respectively, and the difference is marked as a C/G SNP. If, however, the quality of either allele is below 40, the difference is not marked as a SNP. In addition, amosvalidate evaluates the proximity of SNPs to further increase the confidence in the predictions; clusters of SNPs that occur within a small range in the assembly are likely indicative of a mis-assembly. By default, this is defined as regions containing at least 2 high quality SNPs occurring within a 500 bp window.

Note that this technique for mis-assembly detection can also be applied in heterogeneous genomes, for example, by identifying regions with a significantly higher SNP density than the background rate. In such genomes, however, a much higher false-positive rate, due to localized regions of heterogeneity, requires combining this method with other validation measures.

6.2.5 Read Breakpoints

The reads provided to an assembler must be consistent with the resulting assembly. Thus, examining how the un-assembled reads (also called singletons, or shrapnel) disagree with the assembly can reveal potential mis-assemblies. The Nucmer [80,202] alignment program is used to compare un-assembled reads to a consensus, allowing fragmented alignments to the consensus. For instance, a mapping that aligns the first half of a read to a different region than the second half, but at 100% identity, is preferable to a mapping that aligns the read contiguously at 80% identity. The fragmented, high identity alignment is more likely because the read sequence should be nearly identical to the consensus sequence, modulo sequencing errors. From among all alignments of a read to the genome, the placement is chosen that maximizes the sum of $len(A_i) * idy(A_i)$ over all alignment segments A_i, where $len(A_i)$ and $idy(A_i)$ are the length and percent identity of the i^{th} segment of alignment A, and $len(A_i)$ is adjusted where necessary to avoid scoring the overlap between adjacent segments twice. This scoring function estimates the number of non-redundant bases matching the consensus, and the Nucmer utility *delta-filter* computes an optimal alignment using this function and a modified version of the Longest Increasing Subsequence (LIS) algorithm [34]. Most mappings consist of a single alignment that covers the entire read, while the fragmented mappings indicate either incorrect trimming of the read or the presence of a mis-assembly.

For fragmented alignments, the locations where the alignment breaks—boundaries of alignment fragments that do not coincide with the ends of the read—are called "breakpoints". Under the assumption that all reads map perfectly to the assembly, breakpoints indicate the presence of errors, either in the assembly, or in the reads themselves (e.g. incomplete trimming, or chimeric fragments). Breakpoints supported by a single read are rarely cause for concern, and can often be explained by errors in the reads themselves. However, multiple reads that share a common breakpoint often indicate assembly problems. These multiply supported breakpoints are identified, after the alignment process described in the previous section, by sorting the boundaries of fragmented alignments by their location in the consensus, and reporting those that occur in multiple reads. In addition, each read is annotated with a vector of coordinates encoding all breakpoints in the alignment of the read to the genome. This vector helps determine not only if two reads share common breakpoints, but also if they have similar mappings to the consensus. For each breakpoint, the cluster of reads with similar alignment signatures are examined to characterize different classes of mis-assemblies in much the same way mate-pairs are used to characterize collapse, inversion, etc. But while mate-pair and coverage methods can only bound a mis-assembly to a certain region, breakpoints can identify the precise position in the consensus at which the error occurs.

6.2.6 Integrative Validation

The amosvalidate pipeline executes the analyses described above to tag regions that appear mis-assembled. Independently, each analysis method may report many false-positives that reflect violations of the data constraints, but that do not necessarily represent mis-assemblies or incorrect consensus sequence. A common example is clusters of overlapping stretched or compressed mate-pairs caused by a wide variance in fragment sizes rather than mis-assembly. Combining multiple mis-assembly signatures increases the likelihood that the tagged regions identify true errors in the assembly. For example, a region with a largely negative CE value is more likely to indicate the presence of a collapsed repeat if an unusually high density of correlated SNPs is also present. This particular combination is especially strong, since mate-pair and sequence data are independent sources.

Since some types of signatures do not necessarily tag the exact location of a mis-assembly, combining mis-assembly signatures requires considering not only overlapping signatures, but also those that occur in close proximity. To combine mis-assembly signatures, the pipeline identifies regions in the assembly where multiple signatures co-occur within a small window (2 Kbp by default). If multiple signatures of at least two different evidence types occur within this window, the region is flagged as “suspicious”. Each such region is reported along with detailed information about the individual signatures, and forms the initial focus for subsequent validation and correction efforts. For manual analysis, these regions, along with the individual mis-assembly features, can be viewed alongside the assembly data in the AMOS assembly viewer, Hawkeye.

6.2.7 Visualization

Cognitive psychologist and computer science researcher Herbert Simon stated, “Solving a problem simply means representing it so that the solution is obvious” [203]. In this spirit, Hawkeye strives to provide a visual, manipulable interface to help finishers understand and reason about complex assembly data. In addition to providing a useful interface for the examination of assembly data, Hawkeye further supports the analytical process by providing statistical and computational data analysis, enabling users both to reduce data complexity and to form accurate judgments.

Hawkeye addresses the issues of scale and complexity by guiding users to the most likely areas of mis-assembly, and adhering to the visual-information-seeking mantra of: overview first, zoom and filter, then details-on-demand [204]. The main application window, or “Launch Pad”, acts as a global overview by displaying summary assembly statistics, along with graphs and sortable tables of assembly information. The ranking component of this display encourages users to inspect regions of the assembly in order of importance—largest to smallest and low quality to high quality. The more detailed “Scaffold View” is capable of displaying an entire contig or scaffold and its underlying reads on a single screen for scaffolds spanning >10 Mbp of sequence and >100,000 reads (Figure 6.6). Alternatively, users can zoom in and filter the display to focus on particular regions of interest. Finally, the lowest level assembly information is displayed in the coordinated “Contig View”, displaying the consensus sequence, read-tiling, base-calls, and supporting data. Coordination among these three views—Launch Pad, Scaffold View, and Contig View—allows for very efficient top-down analysis of even the largest assemblies, and leads the user to a natural analytic progression: discern high-level quality from statistics and features; examine a poorly scoring scaffold for mis-assembly at the clone-insert level, looking for uneven insert distribution and improperly sized or mis-oriented mate-pairs; examine possible mis-assemblies in more detail at the base-call and raw data level, looking for correlated discrepancies supported by the raw data; and finally, confirm or refute the mis-assembly hypothesis.

6.2.8 Implementation

The validation modules of amosvalidate are implemented in C++ and included as part of the AMOS assembly package (http://amos.sourceforge.net). AMOS is a modular, open-source framework for genome assembly research and development, which provides integration between software modules through a centralized data store and a well defined API. This framework allows developers to focus on a particular area of interest, e.g. scaffolding, without needing to develop a complete assembly infrastructure. Furthermore, AMOS can import data from common assembly programs and formats—ACE, NCBI Assembly/Trace Archives [205], Arachne [206,207], Celera Assembler [182], PCAP [208], Phrap [209], Phusion [210] and Newbler [211], allowing for the integration of AMOS modules into existing assembly pipelines.

6.3 Validation Examples

6.3.1 Tandem Repeat Collapse in *Bacillus anthracis*

The impetus for much of this work was a mis-assembly detected in the parent strain of *Bacillus anthracis Ames Ancestor* (RefSeq ID: NC_007530). As shown in Figure 6.5, an alignment breakpoint analysis detected four unassembled reads which only partially matched the assembly. The partial matches ended at the same locations in all reads, specifically at coordinates 144,337 and 146,944 in the assembled main chromosome of *B. anthracis*. This pattern is consistent with the collapse of a tandem repeat consisting of two copies of the sequence between these two coordinates. The four unassembled reads span the boundary between the two copies of the repeat, leading to the observed alignment in the incorrect assembly. Increased depth of coverage was also observed in the assembly, supporting the collapse hypothesis. This observation was confirmed by a close inspection of the assembly in this region, and the finishing team at TIGR was able to correct the assembly.

It is important to note that this genome had been finished at The Institute for Genomic Research (TIGR) and had already been deposited into GenBank at the time when this mis-assembly was identified. The mis-assembly had thus escaped detection despite the extremely stringent manual curation performed by the finishing teams at TIGR. Since finishing is primarily aimed at closing gaps, rather than fixing mis-assemblies, it is not that surprising that errors persist even in finished data. Examples like this reinforce recent calls for caution when dealing with all assemblies, not just those of draft quality [178].

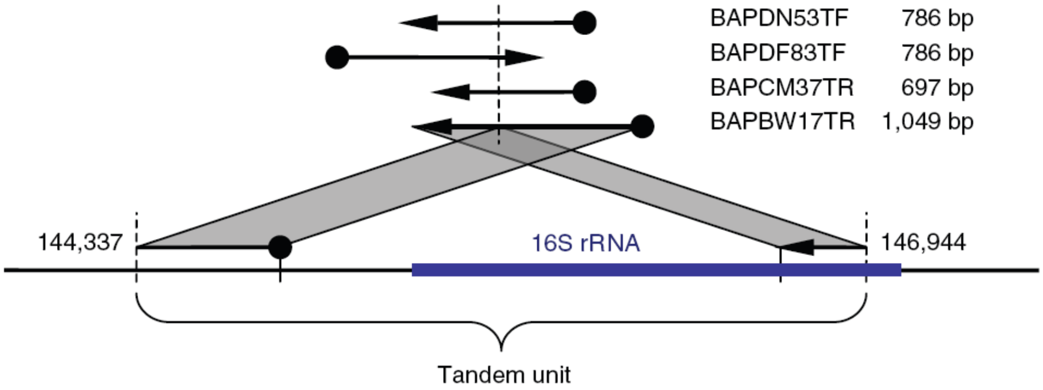


Figure 6.5: *Bacillus anthracis* mis-assembly exhibiting a tandem-collapse breakpoint signature. The alignments of the four reads to the assembly indicate the collapse of a tandem repeat consisting of two copies of the section of the assembly between coordinates 144,337 and 146,944. Note how the alignment signature resembles the mate signature shown in Figure 6.2b.

6.3.2 Mis-assemblies in *Drosophila virilis*

To test the scalability of amosvalidate, the pipeline was run on an assembly of the fruit fly *Drosophila virilis*. The genome was sequenced with the whole-genome shotgun method to ~8x coverage by Agencourt Bioscience Corporation, and assembled with both Celera Assembler and Arachne. The best assembly at the time, Comparative Analysis Freeze 1 (CAF1), comprised 13,530 scaffolds containing 18,402 contigs with a total length of ~189 Mbp. This assembly represents a reconciliation of both the Celera Assembler and Arachne results [195]. Because the read multi-alignment was not provided with the reconciled assembly, this section describes the analysis of a small region of the Celera Assembler assembly. Due to the absence of a finished reference, it is impractical to evaluate the analysis on a larger scale.

In a 556 Kbp contig of the Celera Assembler assembly, amosvalidate predicted 56 mis-assembly signatures and 6 suspicious regions. Two of the suspicious regions are at the extreme ends of the contig, and correctly identify the low quality sequence present at the ends of the contig. Two more regions are weakly supported by CE stretch and missing mate signatures, but do not appear to be egregious mis-assemblies. The remaining two regions, however, reflect obvious mis-assembly. The left-hand region (Figure 6.6), positioned at 78,088–84,132, is supported by alignment breakpoint, missing mate, and correlated SNP signatures. In addition, the cluster of yellow, compressed mates at the bottom of Figure 6.6 correspond exactly with the position of the correlated SNPs. Examination of the multi-alignment at this position reveals two distinct sets of co-assembled reads. This evidence taken together points to a collapse style mis-assembly. The right-hand region (Figure 6.6b), positioned at 89,408–98,979, is more subtle and supported only by CE expansion and SNP signatures. However, the overwhelming severity of the CE expansion caused by the cluster of blue, expanded mates at the bottom of Figure 6.6 suggest that additional sequence has been incorrectly inserted into this region.

The official, reconciled CAF1 assembly does not contain either of these mis-assemblies, independently confirming the amosvalidate analysis. Instead, the suspicious region is broken into multiple contigs, with the left half mapping to contig_16268 of the CAF1 assembly and the right half to contig_16269.

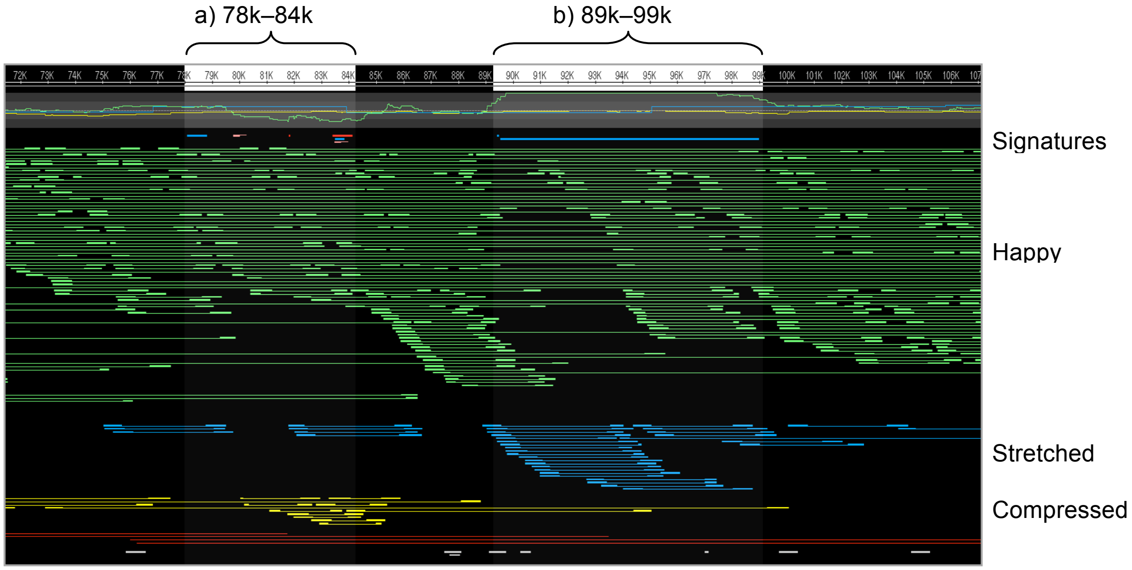


Figure 6.6: Example *Drosophila virilis* mis-assembly shown by Hawkeye. Sequencing reads are represented as thick boxes connected to their mate by thin lines. Correctly sized (happy) mates are shown in green, stretched in blue, and compressed in yellow. A CE statistic plot is given at the top, with mis-assembly signatures plotted directly below as intervals. Left highlight shows *amosvalidate* region (a), which appears to be a compression mis-assembly. Right highlight shows *amosvalidate* region (b), which appears to be an expansion mis-assembly.

6.4 Systematic Evaluation of Bacterial Assemblies

To supplement the anecdotal results presented above, amosvalidate was used to conduct a systematic evaluation of assemblies. Sequencing data for 16 bacterial genomes was collected and assembled with Phrap v0.990329 using the *phrap.manyreads* program with default parameters. Phrap was chosen because of its popularity, simplicity, and tendency to mis-assemble repetitive genomes. Similar experiments were attempted with Celera Assembler, but not enough mis-assemblies were produced to allow adequate validation. In larger genomes, Celera Assembler, and virtually all other assemblers, produce many errors; however, there are not enough fully finished eukaryotic genomes to allow comprehensive testing of automated methods. For extensive and objective testing, bacteria were chosen as the assembly targets because many complete, finished genomes are available, thus providing a proper reference that can be used to identify true mis-assemblies.

The Phrap assemblies were aligned against the reference sequences using the MUMmer utilities *nucmer* and *dnadiff* to collect regions of mis-assembly (http://mummer.sourceforget.net). dnadiff performs a whole-genome alignment and compactly summarizes the location and characteristics of differences between two contig sets. For aligning contigs to a reference genome, this process is identical to the read mapping discussed in the Read breakpoint analysis section. Using the same algorithm, the contig set is mapped to the reference genome using Nucmer, and the optimal mapping for each contig is identified. The alignment information is then parsed, and all alignment breakpoints are identified. By default, Nucmer creates a contiguous alignment as long as the average nucleotide identity is greater than 70% for a 200 bp window; therefore, any stretch of greater than ~60 mis-matches will force the alignment to break. After alignment, the breakpoints are classified as insertions, deletions, rearrangements, or inversions based on their surrounding context. For example, a breakpoint between a forward-strand and negative-strand alignment on the same contig is classified as an inversion. For the Phrap contigs, only alignment differences that produced a breakpoint were considered as mis-assemblies. Small differences such as consensus SNPs, short indels (< ~60 bp), and breakpoints occurring within the first 10 bp of a contig were ignored. All contigs less than 5,000 bp were also ignored because of their generally low quality.

amosvalidate was then run on all 16 Phrap assemblies to determine if the mis-assembled regions were correctly identified. Table 6.1 gives a summary of the Phrap induced mis-assemblies, along with statistics detailing the performance of amosvalidate. Table 6.2 gives specific details on the types of mis-assemblies introduced by Phrap, and the size characteristics of the amosvalidate features. Mis-joins (rearrangements) where the most prevalent type of mis-assembly reported by dnadiff.

In summary, the sensitivity of the methods is quite good. 96.9% of known mis-assemblies are identified by one or more amosvalidate signatures, and 92.6% are identified by one or more amosvalidate suspicious regions. However, the apparent specificity appears quite low. The over-prediction of mis-assembly signatures can be mostly ignored, because each signature represents a true violation of the five rules listed in the introduction. These are meant to highlight inconsistencies in the assembly, and do not always correspond to actual mis-assemblies. The over-prediction of suspicious regions appears to indicate a limitation of automated methods. In this case, it is mostly due to the nature of the Phrap algorithm. Because the version of Phrap used in this analysis disregards mate-pair information, many reads are placed in incorrect repeat copies. This leads to both correlated SNPs in the read multi-alignment and unsatisfied mate-pairs. In some cases, misplacing repetitive reads is benign and the resulting consensus sequence is correct. However, amosvalidate identifies the SNPs and unsatisfied mates as a signature of mis-assembly and reports the region as suspicious. This is arguably the correct behavior, and for the false-positives that were manually investigated, this was indeed the case.

This is also the reason for such a large fraction of some assemblies being marked as suspicious (as high as 50% in some cases, see Table 6.2). Acceptable specificity of the method for more sophisticated assemblers is evidenced by the previous *D. virilis* example, where analysis of the 556 Kbp Celera Assembler contig revealed only 6 suspicious regions that covered 4% of the total sequence.

As would be expected, the wide variance of mis-assemblies found in the Phrap assemblies roughly correlates with genome repeat content, with no mis-assemblies being found in the small, non-repetitive assembly of *Neorickettsia sennetsu*, and 151 being found in the complex assembly of *Xanthomonas oryzae*, which contains many highly repetitive insertion sequence (IS) elements. The quality of these two assemblies is clearly reflected in the percentage of the genome marked as suspicious (3.5% and 55.1% respectively). Also interesting are the 3 mis-assemblies identified in the *Mycoplasma capricolum* assembly, none of which were identified by amosvalidate. Manual inspection of the reference alignment shows tandem repeat expansions of lengths 42, 240, and 654 bp. However, the assembly appears sound at these points with no fluctuation in CE statistic, good coverage, and few unsatisfied mates. Closure teams generally spend extra effort to properly handle repetitive regions, but if these repeats went unidentified during the closure process, it is possible that the reference sequence was mis-assembled. Unfortunately, the original assembly is not available for this genome, and only experimental validation could confirm the exact length and copy number of these repeats.

Table 6.1: Accuracy of amosvalidate mis-assembly detection signatures and suspicious regions summarized for 16 bacterial genomes assembled with Phrap.

				Mis-assembly Signatures			Suspicious Regions		
Species[1]	**Len[2]**	**Ctgs[3]**	**Errs[4]**	**Num[5]**	**Valid[6]**	**Sens[7]**	**Num**	**Valid**	**Sens**
B. anthracis	5.2	87	2	1336	21	100.0	127	2	100.0
B. suis	3.4	120	10	1047	30	80.0	158	9	90.0
C. burnetii	2.0	55	22	1375	70	100.0	124	19	100.0
C. caviae	1.4	270	12	625	16	83.3	50	8	66.7
C. jejuni	1.8	53	5	290	11	80.0	61	3	60.0
D. ethenogenes	1.8	632	12	688	22	91.7	88	9	100.0
F. succinogenes	4.0	455	21	1670	27	95.2	266	14	66.7
L. monocytogenes	2.9	172	1	1381	5	100.0	201	1	100.0
M. capricolum	1.0	17	3	83	0	0.0	16	0	0.0
N. sennetsu	0.9	16	0	91	0	NA	13	0	NA
P. intermedia	2.7	243	21	1655	57	100.0	201	20	100.0
P. syringae	6.4	274	64	2841	200	98.4	366	55	98.4
S. agalactiae	2.1	127	21	687	53	95.2	112	18	85.7
S. aureus	2.8	824	41	1850	69	97.6	227	18	75.6
W. pipientis	3.3	2017	31	761	92	100.0	132	30	100.0
X. oryzae	5.0	50	151	2569	379	100.0	100	69	100.0
Totals	**46.8**	**5412**	**417**	**18949**	**1052**	**96.9**	**2242**	**275**	**92.6**

[1] Species name, [2] genome length, [3] number of assembled contigs, and [4] alignment inferred mis-assemblies. The [5] total count, [6] number coinciding with a known mis-assembly, and [7] percentage of known mis-assemblies identified are given for both mis-assembly signatures and suspicious regions. A signature or region is deemed "validated" if its location interval overlaps a mis-assembled region identified by dnadiff.

Table 6.2: Types of detected mis-assemblies and feature characteristics for the results presented in Table 6.1.

		Mis-assembly types				Mis-assembly Signatures			Suspicious Regions		
Species	**Len**	**Ins[1]**	**Del[2]**	**Join[3]**	**Inv[4]**	**Num[5]**	**aLen[6]**	**%Len[7]**	**Num**	**aLen**	**%Len**
B. anthracis	5.2	0	0	2	0	1336	831	21.5	127	5546	13.6
B. suis	3.4	0	0	7	3	1047	1354	42.2	158	7575	35.6
C. burnetii	2.0	0	0	13	9	1375	1106	74.3	124	11455	69.4
C. caviae	1.4	0	0	11	1	625	320	14.1	50	3896	13.7
C. jejuni	1.8	1	0	3	1	290	613	10.0	61	1981	6.8
D. ethenogenes	1.8	0	0	8	4	688	691	26.5	88	4116	20.2
F. succinogenes	4.0	0	1	19	1	1670	1387	57.5	266	7396	48.8
L. monocytogenes	2.9	0	0	1	0	1381	873	42.1	201	5254	36.9
M. capricolum	1.0	3	0	0	0	83	835	6.8	16	3005	4.7
N. sennetsu	0.9	0	0	0	0	91	512	5.4	13	2328	3.5
P. intermedia	2.7	0	0	19	2	1655	727	44.5	201	6263	46.5
P. syringae	6.4	0	1	43	20	2841	782	34.4	366	5725	32.4
S. agalactiae	2.1	0	0	16	5	687	793	25.6	112	4082	21.5
S. aureus	2.8	1	0	34	6	1850	740	49.0	227	5582	45.4
W. pipientis	3.3	0	0	17	14	761	1206	28.1	132	6395	25.8
X. oryzae	5.0	1	0	74	76	2569	1551	79.0	100	27771	55.1
Totals	**46.8**	**6**	**2**	**267**	**142**	**18949**	**895**	**35.1**	**2242**	**6773**	**30.0**

[1] Tandem insertion, [2] tandem collapse, [3] mis-join, and [4] inversion mis-assemblies. The [5] total count, [6] average length, and [7] total length as a percentage of genome are given for both mis-assembly signatures and suspicious regions.

6.5 Discussion

Due to the high cost of genome finishing, an increasing number of genomes, both prokaryotic and eukaryotic, are sequenced to only a draft level. Efforts at providing quality standards for draft genomes (e.g. the comparative-grade standard [212]) have not yet addressed the issue of large-scale mis-assemblies, leading to the likely possibility that such mis-assemblies are present in the data deposited (at an ever increasing rate) in public databases. In addition, this chapter has shown that mis-assemblies can persist even in "finished" genomes. This situation is particularly troubling as scientists move away from the "gene by gene" paradigm and attempt to understand the global organization of genomes. Without a clear understanding of the errors present in the data, such studies may draw incorrect conclusions. The validation capabilities provided by amosvalidate provide a first step towards a robust set of measures of assembly quality that go beyond the simple base-level measures commonly used. Future work could explore methods for converting mis-assembly features into a type of assembly quality score representing the probability of mis-assembly at any location. The tools presented here, combined with tools designed to correct assemblies, will ultimately lead to automated finishing protocols that could dramatically improve the quality of draft-level assemblies.

6.6 Author Contributions

A version of this chapter appeared previously in published form:

Phillippy AM, Schatz MC, Pop M (2008) Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 9: R55.

This text was coauthored with Michael Schatz and Mihai Pop. I developed the AMOS infrastructure, breakpoint validation, and feature combiner; performed the analysis and generated the results; and drafted the corresponding sections of the text and the introduction. Michael Schatz developed the methods for repeat validation, micro-heterogeneity detection, and visualization; and drafted the corresponding sections of the text. Mihai Pop directed the project; developed the mate-pair validation; performed the *B. anthracis* validation; and drafted the corresponding sections of the text.

Chapter 7

Conclusion

The computational methods and tools presented in this dissertation integrate with advanced biotechnology techniques to create more effective methods for real-time pathogen detection using PCR, comparative genomic analysis using microarrays, and whole-genome sequence assembly. I have contributed novel computational methods enabling more effective design or analysis in each of these three areas. Together they form a comprehensive diagnostic suite capable of both rapidly detecting a pathogen and characterizing its genomic composition at the single nucleotide level. As biotechnology advances, nucleic acid tests will continue to improve in accuracy, cost, and portability. The power and versatility of these tests promise to revolutionize modern diagnostics, leading to more effective pathogen screening and diagnosis, and improving the lives of millions.

In addition, the computational advances in this dissertation are broadly applicable to many other areas of bioinformatics and genome biology. For example, the Insignia database fundamentally represents a distance, in the number and size of matches, between all pairs of sequenced genomes. Because it is quick to compute, relative to other distance metrics, this method could be used for future phylogenetic studies of many whole genomes. Additionally, investigating the nature and function of "signature genes" may help explain what it means to be a bacterial species. To survey environmental diversity, the new pan-genome microarrays introduced here are applicable to metagenomics, and could be used to minimally tile multiple phylogenetic markers for either sequence capture or enrichment from an environmental sample. Finally, improvements in genome assembly and validation not only benefit diagnostics and forensics, but all sequence based analyses. I am in the process of extending these assembly techniques to problematic assembly data sets, such as polyploid eukaryotes and metagenomic samples.

Bioinformatics is a data-driven science, intertwined with rapidly advancing biological experiments that, as this dissertation has shown, are impossible without sophisticated computational support. Advances in technology are leading to data surpluses that must be managed by corresponding advances in bioinformatics. Analyzing this data requires continued contributions from computer scientists in many areas, including algorithms, databases, machine learning, and systems. This dissertation has benefitted greatly from the cross-fertilization between computer science and high-throughput biology, and I plan to continue straddling these areas in order to adapt and invent practical methods that keep advancing genomics and modern diagnostics.

Abbreviations

BAC	bacterial artificial chromosome
bp	base pair (of DNA)
CE-stat	compression-expansion statistic
CGH	comparative genomic hybridization
Ct	threshold cycle
DNA	deoxyribonucleic acid
HG	homologous groups
Kbp	kilo base pairs
LI	*Listeria monocytogenes* lineage I
LII	*Listeria monocytogenes* lineage II
LIII	*Listeria monocytogenes* lineage III
Mbp	mega base pairs
MEM	maximal exact match
PCR	polymerase chain reaction
qPCR	quantitative PCR
ROC	receiver operating characteristic
RNA	ribonucleic acid
SNP	single nucleotide polymorphism
spp.	species
TIGR	The Institute for Genomic Research
VNTR	variable number tandem repeat

Bibliography

1. Urdea M, Penny LA, Olmsted SS, Giovanni MY, Kaspar P, et al. (2006) Requirements for high impact diagnostics in the developing world. Nature 444 Suppl 1: 73-79.

2. Akhras MS (2008) Nucleic Acid Based Pathogen Diagnostics. Stockholm, Sweden: Royal Institute of Technology.

3. Barken KB, Haagensen JA, Tolker-Nielsen T (2007) Advances in nucleic acid-based diagnostics of bacterial infections. Clin Chim Acta 384: 1-11.

4. Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, et al. (2002) Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. Science 296: 2028-2033.

5. Zwick ME, McAfee F, Cutler DJ, Read TD, Ravel J, et al. (2005) Microarray-based resequencing of multiple Bacillus anthracis isolates. Genome Biol 6: R10.

6. Fitch JP, Gardner SN, Kuczmarski TA, Kurtz S, Myers R, et al. (2002) Rapid Development of Nucleic Acid Diagnostics. Proc IEEE 90: 1708-1721.

7. Fitch JP, Raber E, Imbro DR (2003) Technology challenges in responding to biological or chemical attacks in the civilian sector. Science 302: 1350-1354.

8. Willse A, Straub TM, Wunschel SC, Small JA, Call DR, et al. (2004) Quantitative oligonucleotide microarray fingerprinting of Salmonella enterica isolates. Nucleic Acids Res 32: 1848-1856.

9. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, et al. (2002) Microarray-based detection and genotyping of viral pathogens. Proc Natl Acad Sci U S A 99: 15687-15692.

10. Volokhov D, Pomerantsev A, Kivovich V, Rasooly A, Chizhikov V (2004) Identification of Bacillus anthracis by multiprobe microarray hybridization. Diagn Microbiol Infect Dis 49: 163-171.

11. Slezak T, Kuczmarski T, Ott L, Torres C, Medeiros D, et al. (2003) Comparative genomics tools applied to bioterrorism defense. Brief Bioinform 4: 133-149.

12. O'Connell KP, Bucher JR, Anderson PE, Cao CJ, Khan AS, et al. (2006) Real-time fluorogenic reverse transcription-PCR assays for detection of bacteriophage MS2. Appl Environ Microbiol 72: 478-483.

13. Moser MJ, Christensen DR, Norwood D, Prudent JR (2006) Multiplexed detection of anthrax-related toxin genes. J Mol Diagn 8: 89-96.

14. Keim P, Klevytska AM, Price LB, Schupp JM, Zinser G, et al. (1999) Molecular diversity in Bacillus anthracis. J Appl Microbiol 87: 215-217.

15. Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, et al. (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within Bacillus anthracis. J Bacteriol 182: 2928-2936.

16. McBride MT, Masquelier D, Hindson BJ, Makarewicz AJ, Brown S, et al. (2003) Autonomous detection of aerosolized Bacillus anthracis and Yersinia pestis. Anal Chem 75: 5293-5299.

17. Brown K (2004) Biosecurity. Up in the air. Science 305: 1228-1229.

18. Lim DV, Simpson JM, Kearns EA, Kramer MF (2005) Current and developing technologies for monitoring agents of bioterrorism and biowarfare. Clin Microbiol Rev 18: 583-607.

19. Slezak T, Kuczmarski T, Ott L, Torres C, Medeiros D, et al. (2003) Comparative genomics tools applied to bioterrorism defence. Brief Bioinform 4: 133-149.

20. Kaderali L, Schliep A (2002) Selecting signature oligonucleotides to identify organisms using DNA arrays. Bioinformatics 18: 1340-1349.

21. Gordon PM, Sensen CW (2004) Osprey: a comprehensive tool employing novel methods for the design of oligonucleotides for DNA sequencing and microarrays. Nucleic Acids Res 32: e133.

22. Nordberg EK (2005) YODA: selecting signature oligonucleotides. Bioinformatics 21: 1365-1370.

23. Li F, Stormo GD (2001) Selection of optimal DNA oligos for gene expression arrays. Bioinformatics 17: 1067-1076.

24. Tembe W, Zavaljevski N, Bode E, Chase C, Geyer J, et al. (2007) Oligonucleotide fingerprint identification for microarray-based pathogen diagnostic assays. Bioinformatics 23: 5-13.

25. Rahmann S (2003) Fast and sensitive probe selection for DNA chips using jumps in matching statistics. Proc IEEE Comput Soc Bioinform Conf 2: 57-64.

26. Kurtz S (2003) A Time and Space Efficient Algorithm for the Substring Matching Problem. Zentrum für Bioinformatik, Universität Hamburg.

27. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

28. Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K (1995) Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. PCR Methods Appl 4: 357-362.

29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

30. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61-65.

31. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

32. Lay MJ, Wittwer CT (1997) Real-time fluorescence genotyping of factor V Leiden during rapid-cycle PCR. Clin Chem 43: 2262-2267.

33. Chang WI, Lawler EL (1994) Sublinear expected time approximate string matching and biological applications. Algorithmica 12: 327-344.

34. Gusfield D (1997) Algorithms on strings, trees, and sequences: computer science and computational biology. New York: Cambridge University Press.

35. Gardner SN, Lam MW, Mulakken NJ, Torres CL, Smith JR, et al. (2004) Sequencing needs for viral diagnostics. J Clin Microbiol 42: 5472-5476.

36. Currie BJ (2003) Melioidosis: an important cause of pneumonia in residents of and travellers returned from endemic regions. Eur Respir J 22: 542-550.

37. Gilad J (2007) Burkholderia mallei and Burkholderia pseudomallei: the causative micro-organisms of glanders and melioidosis. Recent Pat Antiinfect Drug Discov 2: 233-241.

38. Srinivasan A, Kraus CN, DeShazer D, Becker PM, Dick JD, et al. (2001) Glanders in a military research microbiologist. N Engl J Med 345: 256-258.

39. Zysk G, Splettstosser WD, Neubauer H (2000) A review on melioidosis with special respect on molecular and immunological diagnostic techniques. Clin Lab 46: 119-130.

40. Rotz LD, Khan AS, Lillibridge SR, Ostroff SM, Hughes JM (2002) Public health assessment of potential biological terrorism agents. Emerg Infect Dis 8: 225-230.

41. Gee JE, Sacchi CT, Glass MB, De BK, Weyant RS, et al. (2003) Use of 16S rRNA gene sequencing for rapid identification and differentiation of Burkholderia pseudomallei and B. mallei. J Clin Microbiol 41: 4647-4654.

42. Godoy D, Randle G, Simpson AJ, Aanensen DM, Pitt TL, et al. (2003) Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, Burkholderia pseudomallei and Burkholderia mallei. J Clin Microbiol 41: 2068-2079.

43. Holden MT, Titball RW, Peacock SJ, Cerdeno-Tarraga AM, Atkins T, et al. (2004) Genomic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei. Proc Natl Acad Sci U S A 101: 14240-14245.

44. Nierman WC, DeShazer D, Kim HS, Tettelin H, Nelson KE, et al. (2004) Structural flexibility in the Burkholderia mallei genome. Proc Natl Acad Sci U S A 101: 14246-14251.

45. Tomaso H, Scholz HC, Al Dahouk S, Pitt TL, Treu TM, et al. (2004) Development of 5' nuclease real-time PCR assays for the rapid identification of

the burkholderia mallei//burkholderia pseudomallei complex. Diagn Mol Pathol 13: 247-253.

46. Tomaso H, Pitt TL, Landt O, Al Dahouk S, Scholz HC, et al. (2005) Rapid presumptive identification of Burkholderia pseudomallei with real-time PCR assays using fluorescent hybridization probes. Mol Cell Probes 19: 9-20.

47. Bauernfeind A, Roller C, Meyer D, Jungwirth R, Schneider I (1998) Molecular procedure for rapid detection of Burkholderia mallei and Burkholderia pseudomallei. J Clin Microbiol 36: 2737-2741.

48. Scholz HC, Joseph M, Tomaso H, Al Dahouk S, Witte A, et al. (2006) Detection of the reemerging agent Burkholderia mallei in a recent outbreak of glanders in the United Arab Emirates by a newly developed fliP-based polymerase chain reaction assay. Diagn Microbiol Infect Dis 54: 241-247.

49. Neubauer H, Sprague LD, Joseph M, Tomaso H, Al Dahouk S, et al. (2007) Development and clinical evaluation of a PCR assay targeting the metalloprotease gene (mprA) of B. pseudomallei. Zoonoses Public Health 54: 44-50.

50. Novak RT, Glass MB, Gee JE, Gal D, Mayo MJ, et al. (2006) Development and evaluation of a real-time PCR assay targeting the type III secretion system of Burkholderia pseudomallei. J Clin Microbiol 44: 85-90.

51. Thibault FM, Valade E, Vidal DR (2004) Identification and discrimination of Burkholderia pseudomallei, B. mallei, and B. thailandensis by real-time PCR targeting type III secretion system genes. J Clin Microbiol 42: 5871-5874.

52. U'Ren JM, Van Ert MN, Schupp JM, Easterday WR, Simonson TS, et al. (2005) Use of a real-time PCR TaqMan assay for rapid identification and differentiation of Burkholderia pseudomallei and Burkholderia mallei. J Clin Microbiol 43: 5771-5774.

53. Supaprom C, Wang D, Leelayuwat C, Thaewpia W, Susaengrat W, et al. (2007) Development of real-time PCR assays and evaluation of their potential use for rapid detection of Burkholderia pseudomallei in clinical blood specimens. J Clin Microbiol 45: 2894-2901.

54. Brett PJ, DeShazer D, Woods DE (1998) Burkholderia thailandensis sp. nov., a Burkholderia pseudomallei-like species. Int J Syst Bacteriol 48 Pt 1: 317-320.

55. Gardner SN, Kuczmarski TA, Vitalis EA, Slezak TR (2003) Limitations of TaqMan PCR for detecting divergent viral pathogens illustrated by hepatitis A, B, C, and E viruses and human immunodeficiency virus. J Clin Microbiol 41: 2417-2427.

56. Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, et al. (2005) E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. Genome Biol 6: R78.

57. Vijaya Satya R, Zavaljevski N, Kumar K, Reifman J (2008) A high-throughput pipeline for designing microarray-based pathogen diagnostic assays. BMC Bioinformatics 9: 185.

58. Schriml L, Gussman A, Phillippy K, Angiuoli S, Hari K, et al. (2007) Gemina: A Web-Based Epidemiology and Genomic Metadata System Designed to Identify Infectious Agents. Intelligence and Security Informatics: Biosurveillance: Springer Berlin / Heidelberg. pp. 228-229.

59. Mungall CJ, Emmert DB (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics 23: i337-346.

60. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-470.

61. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20: 207-211.

62. Wang DG, Fan JB, Siao CJ, Berno A, Young P, et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280: 1077-1082.

63. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, et al. (2003) A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med 348: 1953-1966.

64. Volokhov D, Rasooly A, Chumakov K, Chizhikov V (2002) Identification of Listeria species by microarray-based assay. J Clin Microbiol 40: 4720-4728.

65. Doumith M, Cazalet C, Simoes N, Frangeul L, Jacquet C, et al. (2004) New aspects regarding evolution and virulence of Listeria monocytogenes revealed by comparative genomics and DNA arrays. Infect Immun 72: 1072-1083.

66. Call DR, Borucki MK, Besser TE (2003) Mixed-genome microarrays reveal multiple serotype and lineage-specific differences among strains of Listeria monocytogenes. J Clin Microbiol 41: 632-639.

67. Borucki MK, Kim SH, Call DR, Smole SC, Pagotto F (2004) Selective discrimination of Listeria monocytogenes epidemic strains by a mixed-genome DNA microarray compared to discrimination by pulsed-field gel electrophoresis, ribotyping, and multilocus sequence typing. J Clin Microbiol 42: 5270-5276.

68. Zhang C, Zhang M, Ju J, Nietfeldt J, Wise J, et al. (2003) Genome diversification in phylogenetic lineages I and II of Listeria monocytogenes: identification of segments unique to lineage II populations. J Bacteriol 185: 5573-5584.

69. Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, et al. (2005) Applications of DNA tiling arrays for whole-genome analysis. Genomics 85: 1-15.

70. Bertone P, Trifonov V, Rozowsky JS, Schubert F, Emanuelsson O, et al. (2006) Design optimization methods for genomic DNA tiling arrays. Genome Res 16: 271-281.

71. Graf S, Nielsen FG, Kurtz S, Huynen MA, Birney E, et al. (2007) Optimized design and assessment of whole genome tiling arrays. Bioinformatics 23: i195-204.

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

73. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102: 13950-13955.

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

75. Feng S, Tillier ER (2007) A fast and flexible approach to oligonucleotide probe design for genomes and gene families. Bioinformatics 23: 1195-1202.

76. Chung WH, Rhee SK, Wan XF, Bae JW, Quan ZX, et al. (2005) Design of long oligonucleotide probes for functional gene detection in a microbial community. Bioinformatics 21: 4092-4100.

77. Farber JM, Peterkin PI (1991) Listeria monocytogenes, a food-borne pathogen. Microbiol Rev 55: 476-511.

78. Wiedmann M, Bruce JL, Keating C, Johnson AE, McDonough PL, et al. (1997) Ribotypes and virulence gene polymorphisms suggest three distinct Listeria monocytogenes lineages with differences in pathogenic potential. Infect Immun 65: 2707-2716.

79. McNeil LK, Reich C, Aziz RK, Bartels D, Cohoon M, et al. (2007) The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. Nucleic Acids Res 35: D347-353.

80. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.

81. Garey MR, Johnson DS (1979) Computers and Intractability: A Guide to the Theory of NP-Completeness. New York, NY, USA: W. H. Freeman & Co. 338 p.

82. Ausiello G, Protasi M, Marchetti-Spaccamela A, Gambosi G, Crescenzi P, et al. (1999) Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties: Springer-Verlag New York, Inc. 524 p.

83. Johnson D. Approximation algorithms for combinatorial problems; 1973. ACM New York, NY, USA. pp. 38-49.

84. Feige U (1998) A threshold of ln n for approximating set cover. Journal of the ACM (JACM) 45: 634-652.

85. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, et al. (2009) The Listeria transcriptional landscape from saprophytism to virulence. Nature 459: 950-956.

86. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37: D885-890.

87. Team RDC (2008) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

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

89. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557-572.

90. Willenbrock H, Fridlyand J (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics 21: 4084-4091.

91. Pinto FR, Aguiar SI, Melo-Cristino J, Ramirez M (2008) Optimal control and analysis of two-color genomotyping experiments using bacterial multistrain arrays. BMC Genomics 9: 230.

92. Snipen L, Nyquist OL, Solheim M, Aakra A, Nes IF (2009) Improved analysis of bacterial CGH data beyond the log-ratio paradigm. BMC Bioinformatics 10: 91.

93. Snipen L, Repsilber D, Nyquist L, Ziegler A, Aakra A, et al. (2006) Detection of divergent genes in microbial aCGH experiments. BMC Bioinformatics 7: 181.

94. Phillippy AM, Mason JA, Ayanbule K, Sommer DD, Taviani E, et al. (2007) Comprehensive DNA signature discovery and validation. PLoS Comput Biol 3: e98.

95. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, et al. (2007) Multiplex amplification of large sets of human exons. Nat Methods 4: 931-936.

96. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, et al. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4: 907-909.

97. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, et al. (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4: 903-905.

98. Gardan R, Cossart P, Labadie J (2003) Identification of Listeria monocytogenes genes involved in salt and alkaline-pH tolerance. Appl Environ Microbiol 69: 3137-3143.

99. Gardan R, Duche O, Leroy-Setrin S, Labadie J (2003) Role of ctc from Listeria monocytogenes in osmotolerance. Appl Environ Microbiol 69: 154-161.

100. Kathariou S (2002) Listeria monocytogenes virulence and pathogenicity, a food safety perspective. J Food Prot 65: 1811-1829.

101. Roberts AJ, Wiedmann M (2003) Pathogen, host and environmental factors contributing to the pathogenesis of listeriosis. Cell Mol Life Sci 60: 904-918.

102. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, et al. (1999) Food-related illness and death in the United States. Emerg Infect Dis 5: 607-625.

103. Freitag NE, Port GC, Miner MD (2009) Listeria monocytogenes - from saprophyte to intracellular pathogen. Nat Rev Microbiol 7: 623-628.

104. Cossart P (2007) Listeriology (1926-2007): the rise of a model pathogen. Microbes Infect 9: 1143-1146.

105. Rasmussen OF, Skouboe P, Dons L, Rossen L, Olsen JE (1995) Listeria monocytogenes exists in at least three evolutionary lines: evidence from flagellin, invasive associated protein and listeriolysin O genes. Microbiology 141 (Pt 9): 2053-2061.

106. Zhang W, Jayarao BM, Knabel SJ (2004) Multi-virulence-locus sequence typing of Listeria monocytogenes. Appl Environ Microbiol 70: 913-920.

107. Ward TJ, Gorski L, Borucki MK, Mandrell RE, Hutchins J, et al. (2004) Intraspecific phylogeny and lineage group identification based on the prfA virulence gene cluster of Listeria monocytogenes. J Bacteriol 186: 4994-5002.

108. Chen Y, Knabel SJ (2007) Multiplex PCR for simultaneous detection of bacteria of the genus Listeria, Listeria monocytogenes, and major serotypes and epidemic clones of L. monocytogenes. Appl Environ Microbiol 73: 6299-6304.

109. Roberts A, Nightingale K, Jeffers G, Fortes E, Kongo JM, et al. (2006) Genetic and phenotypic characterization of Listeria monocytogenes lineage III. Microbiology 152: 685-693.

110. Meinersmann RJ, Phillips RW, Wiedmann M, Berrang ME (2004) Multilocus sequence typing of Listeria monocytogenes by use of hypervariable genes reveals clonal and recombination histories of three lineages. Appl Environ Microbiol 70: 2193-2203.

111. Liu D, Lawrence ML, Wiedmann M, Gorski L, Mandrell RE, et al. (2006) Listeria monocytogenes subgroups IIIA, IIIB, and IIIC delineate genetically distinct populations with varied pathogenic potential. J Clin Microbiol 44: 4229-4233.

112. De Jesus AJ, Whiting RC (2003) Thermal inactivation, growth, and survival studies of Listeria monocytogenes strains belonging to three distinct genotypic lineages. J Food Prot 66: 1611-1617.

113. Djordjevic D, Wiedmann M, McLandsborough LA (2002) Microtiter plate assay for assessment of Listeria monocytogenes biofilm formation. Appl Environ Microbiol 68: 2950-2958.

114. Pallen MJ, Wren BW (2007) Bacterial pathogenomics. Nature 449: 835-842.

115. Bentley S (2009) Sequencing the species pan-genome. Nat Rev Microbiol 7: 258-259.

116. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, et al. (2007) Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. J Bacteriol 189: 8186-8195.

117. Lefebure T, Stanhope MJ (2007) Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol 8: R71.

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

119. Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, et al. (2008) Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis. Proc Natl Acad Sci U S A 105: 3473-3478.

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

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

122. Bayjanov JR, Wels M, Starrenburg M, van Hylckama Vlieg JE, Siezen RJ, et al. (2009) PanCGH: a genotype-calling algorithm for pangenome CGH data. Bioinformatics 25: 309-314.

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

124. Willenbrock H, Petersen A, Sekse C, Kiil K, Wasteson Y, et al. (2006) Design of a seven-genome Escherichia coli microarray for comparative genomic profiling. J Bacteriol 188: 7713-7721.

125. Castellanos E, Aranaz A, Gould KA, Linedale R, Stevenson K, et al. (2009) Discovery of stable and variable differences in the Mycobacterium avium subsp. paratuberculosis type I, II, and III genomes by pan-genome microarray analysis. Appl Environ Microbiol 75: 676-686.

126. Phillippy AM, Deng X, Zhang W, Salzberg SL (2009) Efficient oligonucleotide probe selection for pan-genomic tiling arrays. BMC Bioinformatics 10: 293.

127. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189.

128. van Dongen S (2000) Graph Clustering by Flow Simulation: University of Utrecht.

129. Doumith M, Buchrieser C, Glaser P, Jacquet C, Martin P (2004) Differentiation of the major Listeria monocytogenes serovars by multiplex PCR. J Clin Microbiol 42: 3819-3822.

130. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, et al. (2001) Comparative genomics of Listeria species. Science 294: 849-852.

131. Barabote RD, Saier MH, Jr. (2005) Comparative genomic analyses of the bacterial phosphotransferase system. Microbiol Mol Biol Rev 69: 608-634.

132. Moralejo P, Egan SM, Hidalgo E, Aguilar J (1993) Sequencing and characterization of a gene cluster encoding the enzymes for L-rhamnose metabolism in Escherichia coli. J Bacteriol 175: 5585-5594.

133. Power J (1967) The L-rhamnose genetic system in Escherichia coli K-12. Genetics 55: 557-568.

134. Hain T, Steinweg C, Kuenne CT, Billion A, Ghai R, et al. (2006) Whole-genome sequence of Listeria welshimeri reveals common steps in genome reduction with Listeria innocua as compared to Listeria monocytogenes. J Bacteriol 188: 7405-7415.

135. Steinweg C, Kuenne CT, Billion A, Mraheil MA, Domann E, et al. Complete genome sequence of Listeria seeligeri, a nonpathogenic member of the genus Listeria. J Bacteriol 192: 1473-1474.

136. Begley M, Gahan CG, Hill C (2002) Bile stress response in Listeria monocytogenes LO28: adaptation, cross-protection, and identification of genetic loci involved in bile resistance. Appl Environ Microbiol 68: 6005-6012.

137. Begley M, Sleator RD, Gahan CG, Hill C (2005) Contribution of three bile-associated loci, bsh, pva, and btlB, to gastrointestinal persistence and bile tolerance of Listeria monocytogenes. Infect Immun 73: 894-904.

138. Cozzani I, Misuri A, Santoni C (1970) Purification and general properties of glutamate decarboxylase from Clostridium perfringens. Biochem J 118: 135-141.

139. Smith DK, Kassam T, Singh B, Elliott JF (1992) Escherichia coli has two homologous glutamate decarboxylase genes that map to distinct loci. J Bacteriol 174: 5820-5826.

140. Waterman SR, Small PL (1996) Identification of sigma S-dependent genes associated with the stationary-phase acid-resistance phenotype of Shigella flexneri. Mol Microbiol 21: 925-940.

141. Cotter PD, Ryan S, Gahan CG, Hill C (2005) Presence of GadD1 glutamate decarboxylase in selected Listeria monocytogenes strains is associated with an ability to grow at low pH. Appl Environ Microbiol 71: 2832-2839.

142. Ryan S, Begley M, Gahan CG, Hill C (2009) Molecular characterization of the arginine deiminase system in Listeria monocytogenes: regulation and role in acid tolerance. Environ Microbiol 11: 432-445.

143. Camejo A, Buchrieser C, Couve E, Carvalho F, Reis O, et al. (2009) In vivo transcriptional profiling of Listeria monocytogenes and mutagenesis identify new virulence factors involved in infection. PLoS Pathog 5: e1000449.

144. Hain T, Hossain H, Chatterjee SS, Machata S, Volk U, et al. (2008) Temporal transcriptomic analysis of the Listeria monocytogenes EGD-e sigmaB regulon. BMC Microbiol 8: 20.

145. Huson DH, Steel M (2004) Phylogenetic trees based on gene content. Bioinformatics 20: 2044-2049.

146. Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21: 255-265.

147. Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, et al. (2004) Phylogenetic discovery bias in Bacillus anthracis using single-nucleotide polymorphisms from whole-genome sequencing. Proc Natl Acad Sci U S A 101: 13536-13541.

148. Orsi RH, Borowsky ML, Lauer P, Young SK, Nusbaum C, et al. (2008) Short-term genome evolution of Listeria monocytogenes in a non-controlled environment. BMC Genomics 9: 539.

149. Nightingale KK, Windham K, Wiedmann M (2005) Evolution and molecular phylogeny of Listeria monocytogenes isolated from human and animal listeriosis cases and foods. J Bacteriol 187: 5537-5551.

150. Nelson KE, Fouts DE, Mongodin EF, Ravel J, DeBoy RT, et al. (2004) Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen Listeria monocytogenes reveal new insights into the core genome components of this species. Nucleic Acids Res 32: 2386-2395.

151. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

152. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.

153. Carter B, Wu G, Woodward MJ, Anjum MF (2008) A process for analysis of microarray comparative genomics hybridisation studies for bacterial genomes. BMC Genomics 9: 53.

154. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254-267.

155. Eisen JA, Fraser CM (2003) Phylogenomics: intersection of evolution and genomics. Science 300: 1706-1707.

156. Cotter PD, Draper LA, Lawton EM, Daly KM, Groeger DS, et al. (2008) Listeriolysin S, a novel peptide haemolysin associated with a subset of lineage I Listeria monocytogenes. PLoS Pathog 4: e1000144.

157. Ragon M, Wirth T, Hollandt F, Lavenir R, Lecuit M, et al. (2008) A new perspective on Listeria monocytogenes evolution. PLoS Pathog 4: e1000146.

158. Orsi RH, Ripoll DR, Yeung M, Nightingale KK, Wiedmann M (2007) Recombination and positive selection contribute to evolution of Listeria monocytogenes inlA. Microbiology 153: 2666-2678.

159. Nightingale KK, Windham K, Martin KE, Yeung M, Wiedmann M (2005) Select Listeria monocytogenes subtypes commonly found in foods carry distinct nonsense mutations in inlA, leading to expression of truncated and secreted internalin A, and are associated with a reduced invasion phenotype for human intestinal epithelial cells. Appl Environ Microbiol 71: 8764-8772.

160. Olier M, Pierre F, Rousseaux S, Lemaitre JP, Rousset A, et al. (2003) Expression of truncated Internalin A is involved in impaired internalization of some Listeria monocytogenes isolates carried asymptomatically by humans. Infect Immun 71: 1217-1224.

161. Glaser P, Rusniok C, Buchrieser C (2007) Listeria Genomics. In: Goldfine H, Shen H, editors. Listeria monocytogenes: Pathogenesis and Host Response. New York, NY: Springer US.

162. Rouquet G, Porcheron G, Barra C, Reperant M, Chanteloup NK, et al. (2009) A metabolic operon in extraintestinal pathogenic Escherichia coli promotes fitness under stressful conditions and invasion of eukaryotic cells. J Bacteriol 191: 4427-4440.

163. Dalton CB, Austin CC, Sobel J, Hayes PS, Bibb WF, et al. (1997) An outbreak of gastroenteritis and fever due to Listeria monocytogenes in milk. N Engl J Med 336: 100-105.

164. Aureli P, Fiorucci GC, Caroli D, Marchiaro G, Novara O, et al. (2000) An outbreak of febrile gastroenteritis associated with corn contaminated by Listeria monocytogenes. N Engl J Med 342: 1236-1241.

165. Ooi ST, Lorber B (2005) Gastroenteritis due to Listeria monocytogenes. Clin Infect Dis 40: 1327-1332.

166. Gahan CG, Hill C (2005) Gastrointestinal phase of Listeria monocytogenes infection. J Appl Microbiol 98: 1345-1353.

167. Sleator RD, Watson D, Hill C, Gahan CG (2009) The interaction between Listeria monocytogenes and the host gastrointestinal tract. Microbiology 155: 2463-2475.

168. Wendlinger G, Loessner MJ, Scherer S (1996) Bacteriophage receptors on Listeria monocytogenes cells are the N-acetylglucosamine and rhamnose

substituents of teichoic acids or the peptidoglycan itself. Microbiology 142 (Pt 4): 985-992.

169. Loessner MJ, Inman RB, Lauer P, Calendar R (2000) Complete nucleotide sequence, molecular analysis and genome structure of bacteriophage A118 of Listeria monocytogenes: implications for phage evolution. Mol Microbiol 35: 324-340.

170. Brussow H, Canchaya C, Hardt WD (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68: 560-602.

171. Casjens S (2003) Prophages and bacterial genomics: what have we learned so far? Mol Microbiol 49: 277-300.

172. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

173. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.

174. Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, et al. (2004) Quality assessment of the human genome sequence. Nature 429: 365-368.

175. She X, Jiang Z, Clark RA, Liu G, Cheng Z, et al. (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. Nature 431: 927-930.

176. Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, et al. (2003) Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol 4: R25.

177. Stein LD (2004) Human genome: end of the beginning. Nature 431: 915-916.

178. Salzberg SL, Yorke JA (2005) Beware of mis-assembled genomes. Bioinformatics 21: 4320-4321.

179. Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2: 231-239.

180. Sutherland GR, Richards RI (1995) Simple tandem DNA repeats and human genetic disease. Proc Natl Acad Sci U S A 92: 3636-3641.

181. Myers EW (1995) Toward Simplifying and Accurately Formulating Fragment Assembly. J Comp Bio 2: 275-290.

182. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, et al. (2000) A whole-genome assembly of Drosophila. Science 287: 2196-2204.

183. Gordon D, Abajian C, Green P (1998) Consed: A Graphical Tool for Sequence Finishing. Genome Research 8: 195-202.

184. Staden R, Beal KF, Bonfield JK (2000) The Staden package, 1998. Methods Mol Biol 132: 115-130.

185. Semple CA, Morris SW, Porteous DJ, Evans KL (2002) Computational comparison of human genomic sequence assemblies for a region of chromosome 4. Genome Res 12: 424-429.

186. Li S, Liao J, Cutler G, Hoey T, Hogenesch JB, et al. (2002) Comparative analysis of human genome assemblies reveals genome-level differences. Genomics 80: 138-139.

187. Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, et al. (2001) A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell 106: 413-415.

188. Istrail S, Sutton GG, Florea L, Halpern AL, Mobarry CM, et al. (2004) Whole-genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci U S A 101: 1916-1921.

189. Huson DH, Halpern AL, Lai Z, Myers EW, Reinert K, et al. Comparing assemblies using fragments and mate-pairs. Lecture Notes in Computer Science; 2001. Springer-Verlag. pp. 294-306.

190. Schatz MC, Phillippy AM, Shneiderman B, Salzberg SL (2007) Hawkeye: an interactive visual analytics tool for genome assemblies. Genome Biol 8: R34.

191. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.

192. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, et al. (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447: 167-177.

193. Bartels D, Kespohl S, Albaum S, Druke T, Goesmann A, et al. (2005) BACCardI--a tool for the validation of genomic assemblies, assisting genome finishing and intergenome comparison. Bioinformatics 21: 853-859.

194. Dew IM, Walenz B, Sutton G (2005) A tool for analyzing mate pairs in assemblies (TAMPA). J Comput Biol 12: 497-513.

195. Zimin AV, Smith DR, Sutton G, Yorke JA (2008) Assembly reconciliation. Bioinformatics 24: 42-45.

196. Arner E, Tammi MT, Tran AN, Kindlund E, Andersson B (2006) DNPTrapper: an assembly editing tool for finishing and analysis of complex repeat regions. BMC Bioinformatics 7: 155.

197. Tammi MT, Arner E, Britton T, Andersson B (2002) Separation of nearly identical repeats in shotgun assemblies using defined nucleotide positions, DNPs. Bioinformatics 18: 379-388.

198. Kim S, Liao L, Tomb JF. A probabilistic approach to sequence assembly validation; 2001. pp. 38-43.

199. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.

200. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186-194.

201. Churchill GA, Waterman MS (1992) The accuracy of DNA sequences: estimating sequence quality. Genomics(San Diego, Calif) 14: 89-98.

202. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30: 2478-2483.

203. Simon HA (1996) The sciences of the artificial (3rd ed.): MIT Press. 231 p.

204. Shneiderman B (1996) The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Proceedings of the 1996 IEEE Symposium on Visual Languages: IEEE Computer Society.

205. Salzberg SL, Church D, DiCuccio M, Yaschenko E, Ostell J (2004) The genome Assembly Archive: a new public resource. PLoS Biol 2: E285.

206. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, et al. (2002) ARACHNE: a whole-genome shotgun assembler. Genome Res 12: 177-189.

207. Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, et al. (2003) Whole-genome sequence assembly for Mammalian genomes: arachne 2. Genome Res 13: 91-96.

208. Huang X, Wang J, Aluru S, Yang SP, Hillier L (2003) PCAP: A Whole-Genome Assembly Program. Genome Res 13: 2164-2170.

209. Green P (1994) PHRAP documentation: ALGORITHMS.

210. Mullikin JC, Ning Z (2003) The phusion assembler. Genome Res 13: 81-90.

211. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

212. Blakesley RW, Hansen NF, Mullikin JC, Thomas PJ, McDowell JC, et al. (2004) An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res 14: 2235-2244.