Characterization of Genomic Islands and Mobile Regions of Microbial Genomes in the Context of Infectious Disease

by

Bhavjinder Kaur Dhillon

B.Sc., University of British Columbia, 2009

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Summer 2016

Approval

Name: **Bhavjinder Kaur Dhillon**

Degree: **Doctor of Philosophy**

Title: ***Characterization of Genomic Islands and Mobile Regions of Microbial Genomes in the Context of Infectious Disease***

Examining Committee: **Chair:** Dr. Dipankar Sen
Professor

Dr. Fiona Brinkman
Senior Supervisor
Professor

Dr. Jack Chen
Supervisor
Professor

Dr. Lisa Craig
Supervisor
Associate Professor

Dr. Margo Moore
Internal Examiner
Professor
Department of Biological Sciences

Dr. Chris Upton
External Examiner
Professor
Department of Biochemistry and Microbiology
University of Victoria

Date Defended/Approved: August 18, 2016

Abstract

Advances in whole genome sequencing (WGS) technologies have created an era in which WGS can routinely be integrated into disease outbreak investigations for the rapid detection and characterization of the causative agents. Although most genomic investigations of outbreaks to date focus on using single nucleotide variations to help track the spread of disease, this dissertation focuses on efforts to improve the characterization of large clusters of horizontally-acquired genes, named genomic islands (GIs), that may cause large phenotypic changes. Such mobile elements contribute a fundamental role in the rapid adaptation of microbial life to various changes in the environment and are known to encode genes involved virulence, antimicrobial resistance (AMR) and alternative metabolism. I present the integration of rich gene annotations of virulence factors (VFs), AMR genes, and pathogen-associated genes into IslandViewer, a web server for the prediction of GIs in addition to the re-design of the web server to now include an interactive genome visualization library named GenomeD3Plot. I also present the application of IslandViewer for GI analysis on real outbreak data from multiple *Listeria monocytogenes* food-borne outbreaks from across Canada to show that isolates from geographically and temporally distinct outbreaks have unique sets of GIs. In addition, I present an analysis coupling the rich AMR gene annotations with GI predictions over a large collection of diverse microbial genera that revealed AMR genes as a whole are not over-represented within GIs, in contrast to VFs as have been previously studied. However, upon breaking down the dataset, certain classes of resistance were found to be associated with such mobile regions. Lastly, I present a WGS study of *L. monocytogenes* to elucidate the contribution of genetic changes to the ability of this pathogen to tolerate and grow in harsh environments, especially cold temperatures, that are important for its role in causing disease. Overall, this work contributes to improved characterization of GIs as well as a better understanding of trends in the role of GIs and mobile regions in the context of AMR and infectious disease.

Keywords: bioinformatics; microbial genomics; genomic islands; horizontal gene transfer; antimicrobial resistance; virulence

For my children.

Reach for the stars,

you can do it all!

Acknowledgements

First, I would like to acknowledge Dr. Fiona Brinkman for the incredible guidance, support, and mentorship throughout the completion of my thesis work. It has been my privilege to learn from such a talented leader in the Bioinformatics community. I would also like to acknowledge my committee members, Dr. Jack Chen and Dr. Lisa Craig for their thoughtful and creative input, discussions and support throughout my studies. To all past and present members of the Brinkman lab, I would not have been able to complete this thesis without a lot of collaboration, discussions and support from all of you. In particular, I would like to acknowledge Matthew Laird, Julie Shay, Claire Bertelli, Geoff Winsor, Morgan Langille, and Shannan Ho Sui for the significant roles they have all played in various aspects of my thesis research. I also owe many thanks to our collaborator Andrew McArthur for the incredible dataset of antimicrobial resistance genes that his group was able to generate for our analyses. I would also like to thank the Public Health Agency of Canada (PHAC) and PulseNet Canada, especially Gary von Domselaar and Aleisha Reimer, for sharing genome sequences and meta-data for the *Listeria monocytogenes* outbreak dataset. I would like to express my thanks to Patricia Hingston and Jessica Chen from the University of British Columbia for their collaboration in the *L. monocytogenes* cold-tolerance study. I would also like to acknowledge SFU's entire MBB department for providing such an inspiring work environment. Finally, I would also like to acknowledge all the funding agencies that have supported my thesis work: CIHR/MSFHR Bioinformatics training program, Genome Canada, Genome BC, NSERC, Alberta Heritage Foundation, and the Simon Fraser University Community Endowment Fund.

To my incredible parents, Jasvinder Singh and Iqbaljit Kaur Khaira, thank you for your guidance and support throughout my many years of education. Harmunpreet Singh, my partner, thank you for all you do to champion my success and support my goals. My daughter Bachan Kaur, thank you for all the joy and happiness you've brought into this journey, and for being so patient with me through it all. My amazing in-laws, Bhawanjit Singh and Baljinder Kaur Dhillon, thank you for the unconditional support you've provided. And to the rest of my family and friends, thank you for always keeping me grounded. I would not have reached this milestone without any of you and will be forever grateful for the important roles you play in my life.

Table of Contents

List of Tables

List of Figures

List of Acronyms

μ_{max}	Maximum Growth Rate
AFLP	Amplified Fragment Length Polymorphism
AMR	Antimicrobial Resistance
ARDB	Antibiotic Resistance Database
ARO	Antibiotic Resistance Ontology
BCCDC	British Columbia Centre for Disease Control
BLAST	Basic Local Alignment Search Tool
CARD	Comprehensive Antibiotic Resistance Database
CFU	Colony Forming Unit
CGS	Core Gene Similarity
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CS	Cold sensitive
CT	Cold tolerant
E-value	Expectation value
GI	Genomic Island
GWAS	Genome-Wide Association Study
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
ICE	Integrative and Conjugative Element
Indel	Insertion/Deletion
INT	Intermediate
IRIDA	Integrated Rapid Infectious Disease Analysis
IS	Insertion Sequence
LRR	Leucine-Rich Repeat
MGE	Mobile Genetic Element
MLS	Macrolide-Lincosamide-Streptogramin
MLST	MultiLocus Sequence Typing
NCBI	National Center for Biotechnology Information
NML	National Microbiology Laboratory
PAI	Pathogenicity Island
PBP	Penicillin Binding Protein

PCR	Polymerase Chain Reaction
PFGE	Pulse Field Gel Electrophoresis
RAPD	Random Amplification of Polymorphic DNA
RBBH	Reciprocal Best BLAST Hit
REI	Resistance Island
RFLP	Restriction Fragment Length Polymorphism
RGI	Resistance Gene Identifier
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
STM	Signature-Tagged Mutagenesis
SVM	Support Vector Machine
T4SS	Type IV Secretion System
TIR	Terminal Inverted Repeat
VCS	Very cold sensitive
VCT	Very cold tolerant
VF	Virulence Factor
VFDB	Virulence Factor Database
VNTR	Variable Number of Tandem Repeats

Chapter 1.

Introduction

Bacteria are the most abundant and ubiquitously distributed organisms on Earth and are found associated with virtually every habitable surface, including on and inside other organisms. In most cases, they are harmless or even provide benefits to their associated hosts, many of which we may not have even discovered. Yet there are mechanisms by which some bacteria have evolved into pathogens and cause infectious diseases, resulting in nearly 10 million deaths annually worldwide (World Health Organization, 2015). Public health agencies across the world employ powerful techniques to isolate pathogens from infected individuals to identify the particular species and strains in order to gain insight into treatment, diagnosis, and surveillance. Epidemiological investigations of outbreaks try to cluster isolates with previously seen isolates to understand how and why an outbreak may have started and how to prevent further spread.

Current bacterial typing methods routinely used in public health laboratories to distinguish between strains of bacteria are not comprehensive as they focus on small proportions of the entire genome (Li et al., 2009; Singh et al., 2006). For example, techniques such as Restriction Fragment Length Polymorphism (RFLP), Amplified Fragment Length Polymorphism (AFLP), and Pulse Field Gel Electrophoresis (PFGE) use restriction enzymes to digest the genome and separate fragments by size on a gel to compare banding patterns. In the case of AFLP, subsets of fragments are amplified using Polymerase Chain Reaction (PCR) before size separation to provide slightly higher resolution of polymorphic regions. These techniques all require a pure culture of the bacteria to be grown before typing. Development of more accurate and rapid PCR-based methods include Random Amplification of DNA (RAPD), Variable Number of Tandem Repeats (VNTR), MultiLocus Sequence Typing (MLST), and Single Nucleotide Polymorphism (SNP) typing, and offer higher resolution because small portions of the

DNA are actually sequenced (Versalovic and Lupski, 2002). DNA microarrays can also be used for typing and are capable of identifying single nucleotide variants (SNVs), but are more expensive. However, the common theme among all of these techniques is that they all under-estimate the actual variation that may be present between isolates and this poses challenges in differentiating between isolates that may appear to be identical using these techniques (Fournier et al., 2007; Li et al., 2009; Wallis et al., 2010). Because these methods do not look at the entire genome, there is a lot of valuable information missing that could facilitate improved characterization of these pathogens. Whole genome sequencing (WGS) can capture high-resolution genotyping data, and now at a comparable cost and speed using next-generation sequencing technologies and has the potential to replace these older molecular typing methods (Aarestrup et al., 2012; Falush, 2009; Schurch et al., 2010).

Sanger sequencing is a first generation sequencing technology. It is highly accurate, but very time consuming and expensive. Next-generation (also referred to as second-generation) sequencing technologies employ methods that drastically reduce the cost and time for sequencing, at the loss of completeness. These include technologies such as Illumina (Solexa) sequencing by synthesis, Roche 454 pyrosequencing, SOLiD sequencing, and Ion Torrent semiconductor sequencing, that all provide useful data but vary in quality and the applications they can support (Loman et al., 2012; Quail et al., 2012). These methods generate much shorter read fragments that must be assembled into contiguous sequences (called contigs). Aligning and assembling short reads from next-generation sequencers into contigs is a complex challenge for the bioinformatics community and dozens of alignment tools have been developed to address this problem (Fonseca et al., 2012). Every alignment tool has particular benefits or pitfalls, and based on many comparisons, no one alignment tool outperforms the others (Hatem et al., 2013; Shang et al., 2014). The choice of alignment method is really dependent on the sequencing platform, the type of organism being sequenced, and the application of the data and should be evaluated by the user. More often than not, gaps remain between contigs of sequence from such technologies and represent regions of the genome that are highly repetitive and hard to sequence or assemble.

Third-generation sequencing technologies, including the single molecule Pacific Biosciences Single Molecule Real Time (SMRT) sequencing and the Oxford Nanopore MinION system, are addressing some of the problems with second-generation technologies by producing longer reads while still being comparable in cost and time, but are still under development. Early reports have shown promise that such technologies can help fill in gaps to produce complete genomes, in conjunction with second-generation technologies or even on their own (Bashir et al., 2012; Koren and Phillippy, 2015). As a result of the development of new sequencing technologies, there has been explosive growth in the number of microbial genomes sequenced and this technique will continue to become an integral component of many microbial analyses, especially for infectious diseases.

1.1. The emerging field of genomic epidemiology

Genomic epidemiology is the use of WGS data in investigations of infectious disease outbreaks (Parkhill and Wren, 2011). The use of whole genomes can provide a very high-resolution view of the single nucleotide variants (SNVs) between isolates in order to track person-to-person spread of a pathogen, or to identify the source of the outbreak accurately. This revolutionary methodology has led to many breakthroughs in epidemiological investigations that would not have been possible using older genotyping methods. A prominent example is that of the 2010 Haiti Cholera outbreak in which WGS was used to accurately identify the source of the outbreak, a result of human-transmission from Nepal as isolates from the two regions only differed by one or two SNVs in the core genome (Orata et al., 2014). Another important study using WGS of clonal lineages of methicillin-resistant *Staphylococcus aureus* revealed clustering of geographically-related isolates and evidence of transmission events between continents, and even detailed the microevolution of this pathogen within a single hospital to reconstruct transmission events that occurred within the hospital or from other sources in the community (Harris et al., 2010). This high-resolution typing is extremely sensitive and significantly improves our ability to track the spread of infectious disease.

Aside from being able to use SNVs between isolates to track the source and spread of outbreaks, SNVs can also be analyzed for functional impact on affected genes

for potential roles in conferring phenotypic changes to the organism. Furthermore, WGS data can be analyzed for larger changes at the gene level generally to evaluate any gene gains or losses that may play a role in the outbreak, such as virulence genes, or to develop better diagnostic tests, drugs or vaccines (Relman, 2011). Microbes very commonly share genetic material horizontally between unrelated organisms (described in more detail in section 1.2 below), so the acquisition of new genes is frequent enough to warrant exploration during outbreak investigations. This information can be important to assess especially in cases where pathogens acquire genes that allow increased transmissibility, toxicity, or resistance to cause the outbreak, rather than seeking an external environmental or social cause. Older studies using PCR-based amplification of known horizontally-acquired resistance genes have demonstrated the spread of, for example, carbapenem-resistance in outbreak associated Acinetobacter (Valenzuela et al., 2007; Zarrilli et al., 2004), or aminoglycoside-resistance in Enterococcus species (Zarrilli et al., 2005), or multi-drug resistance in Enterobacteriaceae (Leverstein-van Hall et al., 2002), and even multi-drug resistance genes between different species (*Enterobacter cloacae* and *Acinetobacter baumannii*) in the same hospital (Naiemi et al., 2005). Having this information on hand would be of utmost importance for effective clinical treatment of infections and could also become routine practice to replace other antimicrobial resistance screening practices. Some recent studies have begun to use WGS to study the role of horizontal gene transfer in outbreaks. For example, studies of *Listeria monocytogenes* outbreaks confirm that closely related isolates differ in content of mobile elements, and these can also differentiate from sporadic cases (Bergholz et al., 2015; Wang et al., 2015). Overall, the use of WGS information for the study of outbreaks and infectious diseases in general is very powerful and will become increasingly conventional in the public health setting.

1.1.1. The IRIDA project

In collaboration with the British Columbia Centre for Disease Control (BCCDC) and the National Microbiology Laboratory (NML), the Brinkman laboratory (Fiona Brinkman is the designated principal investigator) has initiated the Integrated Rapid Infectious Disease Analysis (IRIDA) computational platform to bring WGS into routine practice in diagnostic public health laboratories across Canada. The IRIDA platform will act as a user-friendly

interface for microbiologists and epidemiologists to interpret WGS data by integrating WGS data with traditional microbiological laboratory tests and epidemiological data. Traditionally, WGS data analysis is performed by an expert bioinformaticist who is familiar with the methods of assembling sequence reads into scaffolds and then detecting variants and performing phylogenetic analysis. By standardizing commonly used methods into pipelines, and building interactive visualizations to represent the results of such experiments, WGS can be more easily incorporated into the public health setting without the need for extensive training in bioinformatics methods. For example, the SNVPhyl (Single Nucleotide Variant PHYLogenomics) pipeline in IRIDA aims to automate the steps of taking short read sequences, assembling a genome, calling variants, and generating a phylogenetic tree (http://snvphyl.readthedocs.io/). Furthermore, the phylogenetic relationships can be mapped to geographical locations using an integrated method named GenGIS (Parks et al., 2009). In addition to this, the IRIDA platform will also provide an opportunity for public health agencies to agree on common terminology routinely used in forms and descriptions of outbreaks to build an ontology that has the potential to computationally identify trends or other patterns in outbreak scenarios. Another tool named IslandViewer, developed in the Brinkman lab, will also be integrated into IRIDA for the detection of horizontally acquired clusters of genes that are commonly seen in microbial genomes and will be discussed in more detail in section 1.4.3. This dissertation will also present the latest developments to IslandViewer in Chapter 2 and Chapter 3.

In general, the field of genomic epidemiology is rapidly expanding. Bacterial genomes are relatively small and can be sequenced within a few hours or days (instead of months or years) for thousands of times less than the cost of Sanger sequencing (Loman et al., 2012), so data generation has exploded. However, detailed analysis of microbial genomes has not kept up at a similar pace. There is a need for improved bioinformatics methods for the automated evaluation of genomes, and the IRIDA platform will address some of these shortcomings for the application of genomics to infectious disease outbreak analysis. This dissertation will focus on efforts to improve characterization of mobile elements in microbial genomes, which can be applied in the context of the IRIDA platform as well as for general microbial genomics studies.

1.2. Horizontal gene transfer

A major source of genetic diversity and adaptability of microbial species is contributed to their ability to acquire foreign DNA horizontally from other species, viruses and even eukaryotes and results in acquisition, rearrangement or deletion of genes (Ochman et al., 2000). In fact, the rate at which genes are gained or lost is much higher than the rate of single nucleotide changes for many prokaryotes (Bacteria and Archaea) (Vos et al., 2015). This process generally involves the movement of mobile genetic elements (MGEs) (see section 1.3 below) between unrelated organisms via one of three major mechanisms of horizontal gene transfer (HGT): transformation, conjugation, and transduction.

Transformation is a process mediated by homologous recombination of exogenous double-stranded DNA (dsDNA) that is taken up from the environment (Johnston et al., 2014). One strand of the dsDNA is degraded while the other is internalized and finds a homologous region for integration into the host genome. It was first observed in *Streptococcus pneumoniae* that were able to “transform” into forms with differences in virulence (Griffith, 1928). This process requires the cell to be in a state of competence in which it expresses genes and factors required for the internalization of such naked DNA. Competence is naturally present in a diverse set of taxa from Bacteria and Archaea (Lorenz and Wackernagel, 1994). It can be induced by signaling between cells using peptides or autoinducers by conditions of nutritional starvation and environmental stresses. Transformation could have evolved as a mechanism to allow bacteria to acquire DNA simply for nutrition, genome maintenance, or diversification through acquisition of adaptation genes to survive against natural selection, but this still remains controversial.

Secondly, conjugation is a process by which two cells physically join through a type IV secretion system (T4SS) (Christie, 2001) for the exchange of conjugative plasmids or other integrative and conjugative elements (ICEs) (such as conjugative transposons) from a donor to a recipient cell (Lederberg and Tatum, 1946). A T4SS pilus, sometimes referred to as the sex pilus, is extended from the donor cell to the recipient cell to initiate conjugation. The substrate DNA is pre-processed to become single stranded for transfer

through the T4SS and is then integrated into the recipient genome and replicated upon transfer. Typically ICEs and conjugative plasmids are self-transmissible and encode all machinery for T4SS, integration/excision into/out of cells (Wozniak and Waldor, 2010) (see sections 1.3.2 and 1.3.4 below for more details). Other elements that do not encode the conjugation machinery and are dependent on external conjugation T4SSs are called mobilizable. Of note, conjugation has played a significant role in the spread of AMR genes that are encoded on such plasmids (Bennett, 2008; Courvalin, 1994).

Finally, transduction is the process by which a virus, referred to as bacteriophage (or simply phage), transfers DNA into a host cell. These phages can follow one of two lifestyles upon entering a host: lytic or temperate. Lytic phage directly start replicating upon entry and synthesize viral proteins to produce progeny, while temperate phage can either follow the lytic cycle or integrate into the host genome to become a prophage and replicate with the rest of the host DNA (Guttman et al., 2004). Induction can cause a prophage to enter the lytic cycle either spontaneously or due to environmental changes. Phages are known to be a key source of virulence genes that have played a role in the evolution of several important pathogens (Abedon and Lejeune, 2007; Fortier and Sekulovic, 2013; Penadés et al., 2015). Not only this, but bacterial genetic material can also be transferred by phages through generalized or specialized transduction. Generalized transduction, first described in *Salmonella enterica* serovar Typhimurium (Zinder and Lederberg, 1952) and *Escherichia coli* (Lennox, 1955), is a method by which DNA mis-packaging of bacterial DNA instead of viral DNA occurs during assembly of viral progeny and could potentially transfer any bacterial gene. Specialized transduction occurs when a prophage is induced into the lytic cycle and it imprecisely excises portions of the flanking host DNA that then become integrated into the phage progeny (Hanks et al., 1988).

1.3. Mobile genetic elements

Mobile genetic elements (MGEs) are regions of DNA capable of (1) excising from a host genome, (2) transferring within or between organisms, and (3) integrating into a chromosome at the new location. This can include elements such as prophage, integrons, transposons, and genomic islands (GIs) as will be described in more detail below. MGEs significantly contribute to the extreme genetic diversity seen in microbial life as they can

spread genetic material across traditional species boundaries. Many MGEs are known to carry an array of genes that may offer selective advantages to the recipient cells under various environmental stresses, including virulence, resistance, and alternative metabolism. In this section, I will review the major classes of MGEs, including elements that are capable of integrating into a host genome. Identification of such elements is crucial in the study of microbial genomes in better understanding the genomic features that may differentiate a given organism from other closely related strains. In sections 1.5 and 1.6 I will further describe virulence factors and resistance genes, as these types of genes are known to be shared horizontally between unrelated microbes and would be of utmost importance to evaluate in the framework of infectious diseases.

1.3.1. Prophage

Phage genomes that have integrated into the bacterial host genome are known as prophage. Prophage are very commonly found in diverse bacterial species and can constitute up to 10-20% of the genome (Casjens et al., 2000; Casjens, 2003). They are responsible for a significant portion of the variation present between closely related bacterial strains (Canchaya et al., 2004), as seen in *E. coli* for example (Ohnishi et al., 2001). Prophage are typically integrated into tRNA genes because they contain sequences that are recognized by phage integrases for recombination at that specific location (Campbell et al., 1992; Hacker et al., 1997; Williams, 2002). Most phage carry an integrase, phage-related structural and proliferation genes, as well as an array of other types of genes. Many known toxins and VFs that cause disease are encoded within prophage genes (Banks et al., 2002; Boyd et al., 2001; Cheetham and Katz, 1995; Fortier and Sekulovic, 2013; Penadés et al., 2015; Wagner and Waldor, 2002). For example, the Shiga toxins known to cause severe disease that are produced by some *E. coli* and *Shigella dysenteriae* strains are encoded on prophages and can be spread for the emergence of new Shiga-toxin producing strains (Schmidt, 2001). Similarly, a bacteriophage named CTXϕ carries the cholera toxin gene, which, once integrated into *Vibrio cholerae*, can cause severe disease (Waldor and Mekalanos, 1996). More recently, AMR genes have also been found packaged in prophage molecules (Quiros et al., 2014). Importantly, prophage contribute to the vast genetic diversity of microbial species and have evolved to spread a multitude of different genes via HGT.

1.3.2. Transposons and insertion sequences

Transposons are regions of DNA that are able to move, or transpose, to new locations. Movements are generally within a single genome, but are not limited to this as transposons are often associated with other MGEs such as prophages and GIs that allow HGT into a new genome. Transposons were initially noticed in maize as recurrent breaks in the same region of the genome (McClintock, 1941) that were later attributed to be caused by such “jumping genes”. Since then, transposons have been identified across all domains of life and can vary widely in copy number. They have been found to proliferate to high copy numbers in many plant and animal genomes, contributing to about 50% of the maize genome (SanMiguel et al., 1996) and 35% of the human genome (Smit and Riggs, 1996) for example.

Transposons can vary in molecular structure. Autonomous transposons encode enzymes called transposases for movement to new locations, while non-autonomous transposons do not encode transposases and rely on other transposons for movement. Some transposons are flanked by terminal inverted repeats (TIRs) and have linker sequences between the terminal repeats and the open reading frames. These are called TIR transposons while non-TIR transposons do not carry such a signal. In addition, some transposons may carry a promoter to drive expression of encoded genes. Transposons can carry a variety of different types of genes that could contribute to changes in the host phenotype, including AMR genes such as those found on Tn1403 in *Pseudomonas aeruginosa* (Stokes et al., 2007) and Tn6026 and Tn6029 in *Enterobacteriaceae* (Reid et al., 2015).

Insertion sequences (ISs) are similar to transposons but are generally smaller in size (ranging from 0.7 kb to 2.5 kb) and encode only one or two open reading frames that typically do not contribute to the phenotype of the host (Siguier et al., 2015). ISs may or may not encode a transposase gene and are generally flanked by TIRs. More than 4000 different IS elements have been identified to date and they can be classified into major families based on the transposase sequence (Siguier et al., 2015). Two adjacent IS elements with DNA in between can form a larger mobile element called a composite transposon that can encode multiple genes, including AMR genes. For example, Tn10 is a composite transposon that is flanked by two IS elements and encodes five genes of

which there is one transposase and four potential resistance genes including *tetA* and *tetR* AMR genes against tetracycline (Foster et al., 1981).

The movement of transposons or IS to new locations within and across genomes can have deleterious effects on the host by disrupting genes at insertion sites or altering the expression of neighbouring genes (Mahillon and Chandler, 1998; Siguier et al., 2015). They can also cause chromosomal deletions because of homologous recombination between two transposons and thus they contribute to large-scale reorganization of genomes (Toussaint and Merlin, 2002). However, these elements have also been shown to introduce beneficial mutations, for example, in resting samples of *E. coli* which showed improved growth arising from the movement of IS elements (Naas et al., 1994).

Conjugative transposons, also known as integrative and conjugative elements (ICEs), are transposons that are able to excise to form a circular intermediate that can be shared via conjugation (similar to plasmids) (Alvarez-Martinez and Christie, 2009; Salyers et al., 1995b; Scott, 1992; Wozniak and Waldor, 2010). However, these elements are unlike plasmids in that they are not capable of self-replication and must be integrated into a host genome for replication and survival (Salyers et al., 1995b). Additionally, these elements use different methods for excision and integration than classic transposons that are more similar to bacteriophage, but, since they do not form viral particles, cannot spread via transduction (Salyers et al., 1995b).

In summary, transposons and IS elements are very diverse MGEs that are widely distributed and found at a range of copy numbers in genomes from all domains of life and can contribute to significant genetic diversity.

1.3.3. Integrons

Integrons are essentially a genetic system that utilizes a site-specific recombinase gene (*intI*) to capture exogenous genes (often called cassettes) to integrate into or excise them from the genome at a recombination site (*attI)*. Integrons also include a dedicated promoter (*Pc)* to ensure expression of the captured genes. They were first discovered in the study of the spread of resistance genes (Stokes and Hall, 1989) and were thought to be intrinsically mobile. Since then, two distinct subtypes of integrons have been described:

the mobile integrons that are physically linked to mobile elements for HGT, and the superintegrons that are immobile and capture very large gene cassettes. This section and the remainder of this dissertation will focus on the mobile integrons that are specifically linked to other MGEs such as insertion sequences (ISs), transposons, or conjugative plasmids and are capable of spreading via HGT. Five classes of mobile integrons have been described based on differences in sequences of the integrase genes (Mazel, 2006). These integrons are known to play a large role in the spread of AMR as gene cassettes involved in resistance against all major antibiotic classes have been reported (Recchia and Hall, 1995). Class 1 integrons are commonly seen in clinical isolates and include AMR genes against all known β-lactams, all aminoglycosides, chloramphenicol, and many other resistance classes (Mazel, 2006). But other types of genes with largely unknown functions are also captured by integrons (Boucher et al., 2007; Gillings et al., 2008) and because they are observed in many phylogenetically diverse bacterial species, integrons contribute to significant genetic diversity (Boucher et al., 2007).

1.3.4. Plasmids

Plasmids are any autonomously replicating extrachromosomal genetic elements (Lederberg, 1952; Lederberg, 1998). Many plasmids carry an origin of replication sequence in order to initiate replication, while a few integrate into the chromosome to replicate. Some studies have estimated that nearly half of all known plasmids are conjugative or self-transmissible (encode a T4SS for transfer), and the other half are not, including most of the very large plasmids (Smillie et al., 2010). Non-conjugative plasmids that can still transfer via conjugation using the T4SS of another self-transmissible plasmid are called mobilizable plasmids. Because conjugation can occur between very remotely related organisms (while transduction is more limited because of phage host range), plasmids play a fundamental role in the spread of many genes across microbial life. More importantly, because plasmids are able to self-replicate, there is no requirement for orthologous sequences in the host genome for homologous recombination to integrate the genetic material into the genome, thus the range of plasmids is very broad (Thomas and Nielsen, 2005). Also, plasmids can harbour other types of mobile elements, such as transposons and integrons. Of most interest to the study of infectious disease, plasmids are known to have spread many VFs and AMR genes against antibiotics and heavy metals

between very diverse genera (Carattoli, 2003; Elwell and Shipley, 1980; Kado, 2014; Lanza et al., 2015; Ramirez et al., 2014). For example, the OXA β-lactamases which were shown to have been on plasmids for millions of years, have spread through very diverse phylogenetic groups, including *S. enterica* serovar Typhiumurium*, P. aeruginosa, Ralstonia pickettii, A. baumanii,* and *Legionella gormanii* (Barlow and Hall, 2002). In summary, conjugative and mobilizable plasmids are important vectors for HGT and have contributed significantly to the spread of a variety of genes among microbial life.

1.3.5. Genomic islands

Genomic islands (GIs) are clusters of genes, typically ranging from 8 kb to 200 kb in size (Hacker and Kaper, 2000), of probable horizontal origin in Bacteria and Archaea. The definition of GIs is broad and overlaps with classifications of other MGEs as described above, such as prophages, integrons, conjugative transposons, and ICEs. Figure 1.1 visually depicts the overlap between MGEs and GIs. Using this general term is beneficial for the computational prediction of MGEs integrated into chromosomes, especially in cases where there are no clear detectable features, which could be due to mutations that have significantly altered or destroyed signatures of known transmission or integration mechanisms. Hence, the term GI was coined to generally describe these clusters of horizontally acquired genes and includes several different types of MGEs.

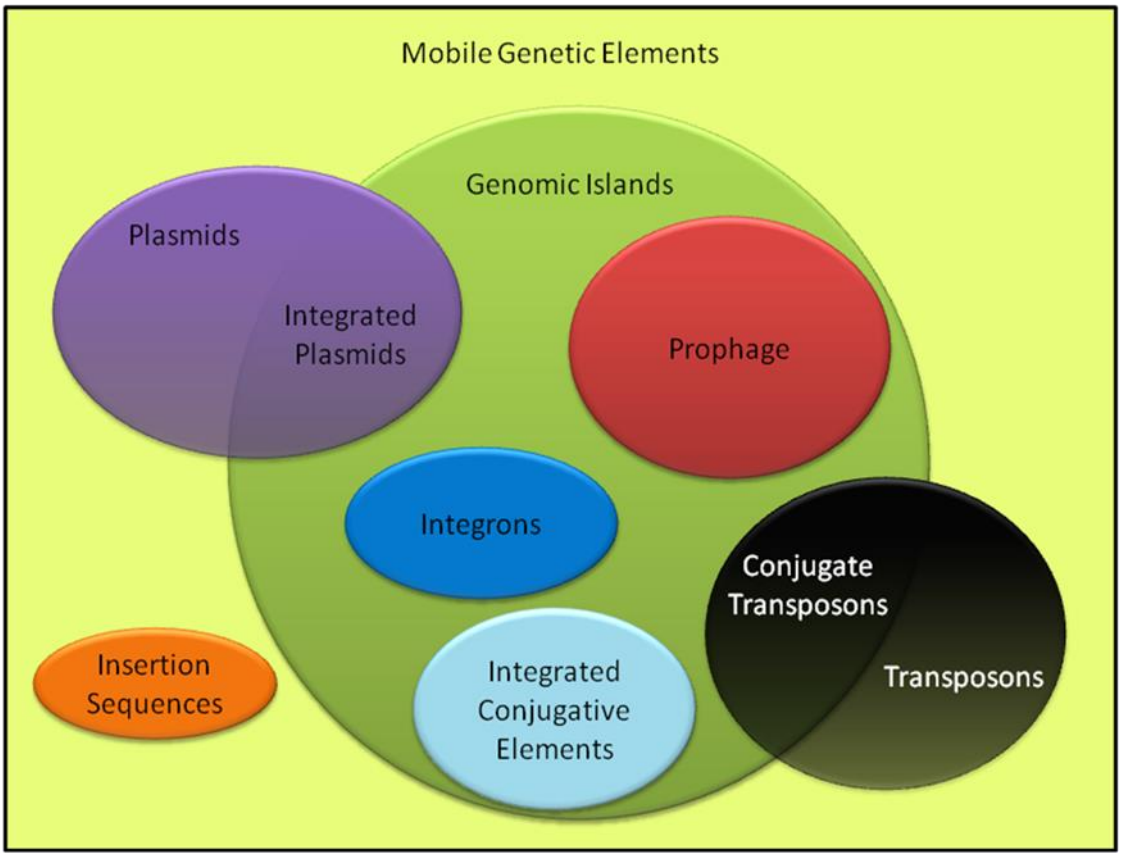

Figure 1.1 Overlap between GIs and other MGEs.
Reprinted from "Computational Prediction and Characterization of Genomic Islands: Insights into Bacterial Pathogenicity" (Doctoral dissertation) by M.G.I. Langille, 2009. Copyright 2009 by M.G.I. Langille. Reprinted with permission.

GIs can be computationally predicted from sequences by examining several different features of such clusters of genes. Various sequence compositional biases exist between host and foreign genomic material. For one, different microbial organisms tend to have varying levels of the four-nucleotide bases in their DNA. Some organisms have high Guanine and Cytosine (GC) content, while other are rich with Adenine and Thymine (AT). Because of this, GIs can be detected by identifying regions of the genome where the GC content differs from the average across the entire genome. However, there could be cases where GC content of foreign DNA is very similar to the host genome and thus this criterion does not capture all GIs. Another more accurate approach is to look at bias in dinucleotides, or codon and amino acid usage as these can also differ between organisms. Other methods can detect bias in *k-mers* of sequence longer than two or three bases, but are known to predict more false positives unless combined with other selective

criteria. In addition to these sequence composition biases, other features of GIs that can be detected include mobility genes (such as transposases and integrases) that are known to be associated with some, but not all, GIs (Hacker et al., 1997). tRNA genes are known insertion sites for phage and are often flanked by direct repeats upon insertion (Hacker et al., 1997), which can also used to improve GI prediction. A machine learning method, Relevance Vector Machine trained on datasets of known GIs and nonGIs (any part of the chromosome not predicted to be a GI) has been previously used to determine the most important identifying structural features of GIs. This study revealed that there exist genus-specific structural families of GIs, but across different genera GIs share a set of core and variable features that include a combination of sequence composition bias, size of the GI, prophages, and presence/absence of integrases as the most important features for GI prediction, while non-coding RNA (including tRNAs), flanking direct repeats, gene density, and insertion point (within coding or intergenic regions) are not as important (Vernikos and Parkhill, 2008). Many computational GI prediction methods identify particular signatures or features of GIs from genome sequences and are discussed in further detail in section 1.4 below.

GIs are extremely important to identify in the study of microbial genomes, especially infectious disease pathogens, since they can spread a variety of genes in one event between unrelated microbes to help them adapt and survive to changing environments (Dobrindt et al., 2004; Hacker and Carniel, 2001). The first identified GIs were termed pathogenicity islands (PAIs) because they encoded a cluster of genes that conferred changes in pathogen virulence of *E. coli* (Hacker et al., 1990). Following this, many other studies discovered that not only VFs were associated with such horizontally acquired regions, but other types of genes like alternative metabolism, symbiosis, resistance, and adaptation in general were found on similar islands (Dobrindt et al., 2004; Groisman and Ochman, 1996; Hacker et al., 1997; Hacker and Kaper, 2000; Schmidt and Hensel, 2004). For example, islands encoding genes to increase metabolic capabilities such as sucrose uptake can be found in *Salmonella senftenberg* (Hochhut et al., 1997), or nitrogen fixation in *Wolinella succinogenes* (Baar et al., 2003) and can help these organisms survive environmental changes that may have otherwise limited their growth.

It is clear that GIs can carry numerous types of genes, but there are certain types of genes that have been shown to be disproportionately found on GIs than the remaining chromosome (nonGIs). For instance, previous research in Dr. Fiona Brinkman's laboratory has provided evidence that GIs are a primary source of novel genes for bacteria (Hsiao et al., 2005). This was discovered because a large number of hypothetical proteins of unknown function and with no known homologs were found to be associated with GIs. These are thought to originate from the phage gene pool, which is much larger than the bacterial gene pool (Suttle, 2005), and represents a source of greater diversity for bacteria. Other studies have shown phage related genes in general are associated with GIs (Vernikos and Parkhill, 2008; Waack et al., 2006). Clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated genes (Cas) have also been found to be associated with GIs (Ho Sui et al., 2009) and are known to be carried by various plasmids, megaplasmids, and even phage (Sorek et al., 2008). CRISPR/Cas genes provide "immunity" against incorporation of additional horizontally acquired genetic material and are thought to have developed as a method for phage to prevent further infection by other viruses and plasmids into a genome (Horvath and Barrangou, 2010).

Furthermore, VFs are known to be significantly associated with GIs (Ho Sui et al., 2009). Not only this, but genera of pathogens capable of adapting to different niches had higher proportions of VFs in GIs, while pathogens with more limited lifestyles and HGT, such as intracellular pathogens, had similar levels of VFs inside and outside of GIs. Thus, GIs play an extremely important role in the evolution of virulence in bacterial pathogens. Moreover, there are certain types of VFs that are found at higher levels on GIs: type III and type IV secretion systems and their effector proteins, toxins, adherence factors, capsule formation, iron uptake, and antiphagocytosis VFs. Further classification of VFs as playing "offensive", "defensive", "nonspecific" or "regulatory" roles showed that "offensive" VFs which are involved in invasion of a host or directly causing damage to a host are very significantly associated with GIs followed by "nonspecific" VFs that are either both offensive and defensive, or neither. The association of VFs with GIs also indicates that the evolutionary pressures presented to VFs, especially "offensive" VFs, select for these VFs to remain mobile rather than becoming stable parts of an organisms' genome mainly because they may not always provide an advantage to the microbe. Some VFs may actually be detrimental and damage or kill the host (e.g. toxin VFs that destroy the host

cell), which may only be beneficial in certain niches. So from an evolutionary perspective, it is beneficial for such VFs to remain mobile.

In addition, several AMR genes are known to be spread via GIs and plasmids as described earlier. However, no study has ever examined whether AMR genes are disproportionately associated with mobile elements in comparison to the rest of the genome. Chapter 4 of this dissertation focuses on investigating this particular topic in greater detail.

On top of encoding additional genes, GIs can insert into coding regions of the chromosome to alter or disrupt the expression of host genes. In *Streptococcus pyogenes*, for example, GIs have been shown to induce a mutator phenotype during stationary phase by integrating into and silencing the DNA mismatch repair gene *mutL*, but this phenotype is reversed upon excision of the GI during exponential growth phase (Nguyen and McShan, 2014). This helps the organism balance between acquiring deleterious mutations and diversification that may help adapt to changing environments. Another observed phenomenon is the disruption of competence machinery across multiple species of *S. pneumoniae* upon acquisition of a GI that inhibits transformation and would limit the acquisition of additional horizontally acquired genetic material, similar to the CRISPR/Cas immunity system (Croucher et al., 2016).

In all, GIs have been previously shown to be associated with a variety of genes, including phage, VFs, CRISPR/Cas systems, and a large collection of uncharacterized or novel genes, all of which play significant roles in the evolution of bacterial species. Furthermore, GIs have also been shown to carry resistance genes, alternative metabolism genes and other adaptation genes that help microbes adjust to environmental changes. They can also disrupt or alter the expression of existing host genes. From an infectious disease perspective, GIs may play a role in changing microbial phenotypes rapidly, for example by acquiring a set of virulence genes that suddenly turn a harmless strain into a pathogenic strain, or by allowing microbes to survive through sanitation procedures by acquiring genes for resistance against sanitation. All of these reasons contribute to a general interest in identifying GIs, in addition to SNVs, in newly sequenced microbial genomes, especially in the context of infectious diseases and outbreaks.

1.4. Computational prediction of genomic islands

This section will focus on describing the methods used for computational prediction of GIs, some of which have been previously evaluated for accuracy using a positive dataset of GIs (Langille et al., 2008). Below, I will present the accuracy for any methods as calculated in this previous study and comment on the accuracy of newer algorithms based on their own separate evaluations. A summary of the various algorithms developed to date, their availability and most recent references is presented in Table 1.1. A recent review by Lu and Leong also summarizes the current state of available methods for GI prediction (Lu and Leong, 2016a). Overall, GI detection from genome sequences can be done using one of two approaches: sequence composition-based or comparative genomics-based algorithms.

Table 1.1 Availability of various GI prediction programs

Program Name	Availability	Link/URL	Most recent reference
Alien_Hunter / IVOM	Command line tool	http://www.sanger.ac.uk/science/tools/alien-hunter	(Vernikos and Parkhill, 2006)
Centroid	Command line tool	Available on request	(Rajan et al., 2007)
CGS	Command line tool	Available on request	(Elhai et al., 2012)
DarkHorse	Web resource and source code available	http://darkhorse.ucsd.edu/	(Podell et al., 2008)
EGID	Standalone Java tool	http://www5.esu.edu/cpsc/bioinfo/software/EGID	(Che et al., 2011)
GC Profile / Z-island	Web resource and source code available	http://cefg.uestc.edu.cn/Zisland_Explorer	(Wei et al., 2016)
GIHunter	Web resource and source code available	http://www5.esu.edu/cpsc/bioinfo/dgi/index.php	(Wang et al., 2011)
GI-POP	Web resource	http://gipop.life.nthu.edu.tw/	(Lee et al., 2013)
GIPSy	Standalone Java tool	http://www.bioinformatics.org/groups/?group_id=1180	(Soares et al., 2015)
GIST	Standalone Java tool	http://www5.esu.edu/cpsc/bioinfo/software/GIST	(Hasan et al., 2012)
GI-SVM	Command line tool	https://github.com/icelu/GI_Prediction	(Lu and Leong, 2016b)

Program Name	Availability	Link/URL	Most recent reference
HGT-DB	Web resource	http://genomes.urv.es/HGT-DB/	(Garcia-Vallve et al., 2003)
IGIPT	Web resource	http://bioinf.iiit.ac.in/IGIPT/	(Jain et al., 2011)
INDeGenIUS	Command line tool	Available on request	(Shrivastava et al., 2010)
Islander	Web resource and source code available	http://bioinformatics.sandia.gov/islander/	(Hudson et al., 2015)
IslandPath-DIMOB	Command line tool	http://www.pathogenomics.sfu.ca/islandviewer	(Langille et al., 2008)
IslandPick	Web resource and source code available	http://www.pathogenomics.sfu.ca/islandviewer	(Langille et al., 2008)
IslandViewer	Web resource and source code available	http://www.pathogenomics.sfu.ca/islandviewer	(Dhillon et al., 2015)
MGSIP	Standalone Java tool and source code available	http://msgip.integrativebioinformatics.me	(de Brito et al., 2016)
MobilomeFinder	Web resource	http://db-mml.sjtu.edu.cn/MobilomeFINDER/	(Ou et al., 2007)
PAIDB	Web resource	http://www.paidb.re.kr/	(Yoon et al., 2015)
PAI-IDA	Command line tool	http://omictools.com/pai-ida-tool	(Tu and Ding, 2003)
PIPs	Standalone Java tool	http://www.genoma.ufpa.br/lgcm/pips	(Soares et al., 2012)
PredictBias	Web resource	http://www.bioinformatics.org/sachbinfo/predictbias.html	(Pundhir et al., 2008)
SIGI-HMM	Command line tool	http://www.uni-goettingen.de/en/research/185810.html	(Waack et al., 2006)

1.4.1. Sequence-Composition Approaches

As described previously in section 1.3.5, GIs can be detected based on many distinctive sequence features. Table 1.2 summarizes the various sequence features and types of genes found in GIs that are detected by existing methods, and the algorithms that detect each. Some algorithms focus on a detecting a single distinguishing feature. For example, **SIGI-HMM** detects codon usage bias (Waack et al., 2006) and is one of the most accurate methods (86% accuracy) (Langille et al., 2008). **MGSIP** is a recently developed

method for detection of GC bias without the use of a sliding window, which is the most common approach for older methods (de Brito et al., 2016). Other methods can detect bias in *k*-mers of sequence longer than two or three bases and include tools like **Alien_Hunter** (Vernikos and Parkhill, 2006) (70% accuracy), **Centroid** (Rajan et al., 2007) (82% accuracy), but because they have been shown to be less accurate they were not included in our GI analyses. **INDeGenIUS** (Shrivastava et al., 2010) also detects *k*-mer bias, but the evaluation was performed on a simulated GI dataset and is not comparable to other accuracy measurements. **Islander** detects a tRNA and a fragment of tRNA and considers the region between the two as a GI (Hudson et al., 2015; Mantri, Williams 2004).

On the other hand, most GI prediction methods detect one or more sequence compositional biases in combination with other structural features of GIs to reduce false predictions. **IslandPath-DIMOB**, which identifies dinucleotide biases and mobility genes as the signature for GIs (Hsiao et al., 2005; Hsiao et al., 2003; Langille et al., 2008), has been shown to also be one of the most accurate methods with 86% accuracy like SIGI-HMM (Langille et al., 2008). **PAI-IDA** (Tu and Ding, 2003) and the method used by **HGT-DB** (Garcia-Vallve et al., 2000) identifies GC content, dinucleotide bias and codon usage bias, but has decreased accuracy (83%) (Langille et al., 2008) when compared to other methods. Integrated Genomic Island Prediction Tool (**IGIPT**): combines GC content bias, dinucleotide bias, codon bias, amino acid bias, and larger *k-mer* bias, but an evaluation of accuracy was not performed by the authors (Jain et al., 2011). **GIHunter** also integrates multiple features including sequence composition bias, mobility genes, tRNA genes, phage information, gene density, intergenic distance, and highly expressed genes, in addition to merging clustered GIs (Wang et al., 2011). However, a clear accuracy assessment is not provided for this method as the merging process may result in non-GI genes being falsely included in GIs. There are cases where extremely large GIs have been detected using GIHunter that may actually be multiple smaller GI insertion events that should not be clustered into a large GI. The Core Gene Similarity (**CGS**) method calculates differences in the least frequent *k-mers* of each protein coding gene against the average across “core” genes, or those genes that have orthologs against a set of other bacterial genomes, to pick out genes with foreign signatures (Elhai et al., 2012). The most recently reported **Z-island** method detects GC, codon usage, and amino acid usage biases without using a sliding window technique and is reported to have 91% overall accuracy, however

this number represents an average accuracy of GI prediction in a limited dataset of only 11 genomes and requires further evaluation (Wei et al., 2016).

Machine learning algorithms also exist that include a step of training on collections of known GIs and nonGIs. **GI-POP** is developed as a GI prediction method that is part of an annotation pipeline in order to predict GIs in draft genomes (Lee et al., 2013). The annotation service predicts coding and non-coding genes, searches for Clusters of Orthologous Groups (COGs), and then performs GI prediction using Genome Profile Scanning (GI-GPS), which includes a Support Vector Machine (SVM) classifier trained on known GIs and nonGIs, identification of mobility genes and refinement of GI boundaries based on location of probable tRNAs and repeating elements. However, I have been unable to access this web server since the publication of this method despite multiple attempts and did not include this tool in any analyses. **GI-SVM** is another machine learning based approach that is trained on a set of known GI and nonGI genes (Lu and Leong, 2016b).

There are also tools that attempt to further classify GIs as PAIs or resistance islands (REIs) based on the presence of genes involved in virulence or resistance, respectively. **PredictBias** detects GC content, dinucleotide bias and codon usage bias similar to PAI-IDA, but also further classifies islands as PAIs if any genes within the GI have sequence similarity against a collection of known VF proteins (Pundhir et al., 2008). **PAIDB** also identifies GC content bias and codon usage bias in addition to sequence similarity against known PAI and REI genes (Yoon et al., 2005). PAIDB v2.0 improves on GI prediction by incorporating SIGI-HMM and IslandPath-DIMOB (Yoon et al., 2015). A barcoding method was used to detect PAIs in *E. coli* O157:H7 where *k-mer* frequencies and their reverse complement frequencies were calculated as genomic barcodes (Wang et al., 2010), but may not specifically identify only PAIs. Another standalone tool, **PIPs**, identifies PAIs by detecting GC and codon usage bias, the presence of specific types of genes (i.e. VFs, hypothetical proteins, transposases, flanking tRNAs) and that the island is absent from non-pathogenic organisms (Soares et al., 2012). **GIPSy**, developed by the same authors, expands PIPs to also classify REIs and metabolic islands by searching against databases of resistance and metabolic genes (Soares et al., 2015).

In all, there exist many different algorithms for the prediction of GIs based solely on sequence compositional biases and structural features. But, as with any predictive algorithms, there are cases where sequence-composition based approaches do not work well. For example, false negatives may be present due to the amelioration of ancient GIs into the genome (Lawrence and Ochman, 1997) or because of highly expressed genes, such as genes within ribosomal protein operons, that inherently have different bias in sequence composition from the rest of the genome (Karlin et al., 1998). Because of these reasons, comparative genomics methods can be used to detect GIs that can be missed using the above outlined techniques.

Table 1.2 GI features detected by sequence composition-based approaches

Method	GC bias	Dinuc. bias	Codon usage bias	Amino acid bias	Larger *k-mer* bias	SVM classifier	Mobility genes	tRNA genes	Phage genes	Hypo-thetical genes	Repeats	Gene density	Gene position
Alien_Hunter					X								
Centroid					X								
CGS					X								
GC Profile / Z-island	X		X	X									
GIHunter	X	X	X				X	X	X			X	
GI-POP						X	X	X			X		
GI-SVM					X	X							
HGT-DB	X		X	X									X
IGIPT	X	X	X		X								
INDeGenIUS					X								
Islander								X					
IslandPath-DIMOB		X					X						
MGSIP	X												
PAIDB	X		X										
PAI-IDA	X	X	X										
PIPs	X		X				X	X		X			
PredictBias	X	X	X				X						
SIGI-HMM			X										

1.4.2. Comparative Genomics Approaches

Since GIs are mobile and unstable, they can be sporadically distributed among even the closest phylogenetically related strains. Thus, an alternative approach to predicting GIs is to compare your genome of interest against several closely related genomes of varying phylogenetic distances, to identify regions that are unique and missing from the other genomes. This method can detect very recently acquired GIs or older GIs, depending on which comparative genomes are selected, very closely related strains to detect the former or more divergent strains to detect the latter. Thus, this method heavily relies on the distances of the selected genomes. For instance, if genomes are too closely related, any GIs that inserted before the divergence of the genomes will be missed. On the other hand, if very distant genomes are selected, there may be genome rearrangements present that make alignment difficult and could lead to false positive predictions. In addition to this, some genomes may not have appropriate comparative genomes available to detect GIs at all. But this issue is diminishing as genomes are continually being sequenced and most well-studied species have been extensively sequenced already. Nonetheless, because of the limitations with comparative genomics approaches, a combination of both types of methods would be ideal for GI prediction.

MobilomeFinder is one comparative genomics tool that compares several genomes to identify unique regions in your genome of interest as potential GIs (Ou et al., 2007). It also requires the presence of a tRNA gene nearby the GI, which makes the method quite robust, but limits GI predictions for cases where tRNAs are not used as insertion sites. Another limitation of this method is that it requires manual selection of comparison genomes. For one matter, this makes the method unsuitable for automated analyses. Additionally, this can present inconsistencies in the selection of comparison genomes as species have widely varying phylogenetic distances within different genera that users may be unaware of and would ultimately have significant impact on the final GI predictions.

DarkHorse is another method for performing more comparative analyses for GI prediction that uses a phylogenetic approach comparing predicted lineages for each protein sequence (Podell and Gaasterland, 2007). It uses BLAST to search every protein

against a database of proteins and taxonomy information to calculate a lineage probability index (LPI), which is higher for closely related lineages and lowest for distant ones. The LPI is then used to rank protein for probability of being horizontally acquired. Inherently, this method would not be able to detect novel sequences acquired via HGT that are not present in the database, and this would also lead to false positive predictions.

An alternative method, **IslandPick**, was developed by Langille *et al.* as the most accurate comparative genomics method for identifying GIs (Langille et al., 2008). It relies on a database of complete sequenced genomes available from the National Center for Biotechnology Information (NCBI), called MicrobeDB (Langille et al., 2012) that is further annotated to include computed phylogenetic distances between each genome. These distances are calculated using CVTree (Qi et al., 2004; Xu and Hao, 2009) and are then used to automatically select a set of 4-6 genomes to compare against to identify regions that are unique (i.e. GIs) in your genome of interest. By using CVTree distances, the automated selection of suitable genomes within particular distances ensures consistent performance of IslandPick between various genera. Additionally, this method can be run any number of times using manual selections of comparison genomes, which is important if researchers would like to include particular strains for comparison. Following selection of comparison genes, pairwise genome alignments are performed between the query genome and each selected comparison genome using Mauve (Darling et al., 2004). Any unique regions larger than 8 kb present in the query genome that cannot be aligned against any comparison genome are collected. An additional filtering step is then used to remove unique regions that may actually be the result of duplication events that are not aligned by Mauve because it enforces one-to-one alignment. Each identified unique region is used as input for a BLAST search against the query and all comparison genomes. If the unique region contains BLAST hits that cover more than 4 kb, it is removed and all remaining unique regions are considered putative GIs. This step increases the precision of IslandPick, and overall this method shows the most agreement with curated datasets of known GIs and is used as a gold-standard method for GI prediction. It has also previously been used to build a positive dataset of GIs to evaluate the accuracy of other methods as mentioned in the previous section (Langille et al., 2008).

1.4.3. Combinatorial methods

IslandViewer

IslandViewer (Dhillon et al., 2013; Dhillon et al., 2015; Langille and Brinkman, 2009), a web server used for the prediction of GIs in microbial species, incorporates three of the most accurate GI prediction methods: IslandPath-DIMOB, SIGI-HMM, and IslandPick. None of these methods alone can predict all GIs accurately and each method often provides slightly different results, so this combinatorial approach allows the identification of a variety of complementary GI predictions. IslandViewer results are simply the union of all predictions from the three algorithms. Existing algorithms for GI prediction at the time IslandViewer was initially developed were evaluated based on availability of downloadable software, automation of methods, and high specificity and accuracy against IslandPick. Both IslandPath-DIMOB and SIGI-HMM met all the selection criteria and were reported to have the highest specificity (86-92%) and overall accuracy (86%) and were chosen for incorporation into IslandViewer. Methods with very low specificity were not considered in order to reduce the number of false predictions within IslandViewer.

The web server hosts pre-computed GI predictions for all complete microbial genomes from the NCBI database and is updated regularly, but also allows custom analyses of user-uploaded genomes. The user-friendly interface provides visualizations of results with an overview of the union of GI predictions from all methods, as well as predictions broken down by method. Genome annotations are also available for exploring the types of genes found within predicted GI regions. All results can also be downloaded in various formats, including FASTA and Genbank files that can be used as input for other tools such as Artemis Comparison Tool (ACT) (Carver et al., 2005).

IslandViewer has a large user base and is widely used for the identification of GIs in a variety of different studies. This is evident through the submission of hundreds of custom genomes every month for analysis through the IslandViewer pipeline and the 362 citations of the IslandViewer papers currently as calculated by Google Scholar. This dissertation will present significant improvements that have been made to IslandViewer since its original release in 2008, which includes a major rebuild of genome visualizations

and expanded genome annotations for better study of GI regions, as well as the rest of the genome.

Other methods

Ensemble algorithm for Genomic Island Detection (**EGID**), includes 5 tools for GI prediction: AlienHunter, IslandPath, SIGI-HMM, INDeGenIUS, and PAI-IDA (Che et al., 2011). From previous evaluations, it is known that methods like AlienHunter and PAI-IDA have very low specificity and can produce many false positive predictions. But the EGID ensemble algorithm collectively assesses overlap between GI predictions from all programs to provide a score that demonstrates the level of support for each GI region to balance between sensitivity and specificity. This method has been developed as a standalone Java tool and does not include any pre-computed results for reference genomes. To improve the user-interface of this tool, **GIST** was developed more recently (Hasan et al., 2012). The accuracy of EGID was assessed against an IslandPick dataset, but since IslandViewer inherently incorporates the IslandPick method, EGID is not as accurate as IslandViewer.

1.4.4. Recent developments in improved web-based visualizations

Some of the GI prediction methods presented in this section also generate a variety of visualizations for the interpretation of predictions. Generally, these are static images that show the position of GIs along the genome. There is a need to integrate GI predictions with better general genome browsers for navigating through other genomic annotations and features in the context of these predictions. In the case of web-based GI prediction tools, the powerful capabilities of modern browsers using the latest version of Hyper Text Markup Language (HTML5) can be manipulated to generate dynamic and interactive visualizations. There are javascript libraries specifically designed to handle data-rich visualizations in a more dynamic and flexible way, such as D3 (https://d3js.org/), Chart.js (http://www.chartjs.org/) or dygraphs (http://dygraphs.com/), that could be manipulated for the visualization of genomic data. Other libraries have also been developed specifically for genomic data such as iobio (http://iobio.io/), including bam.iobio (Miller et al., 2014), that is capable of streaming data from large data files of various genome analysis formats for rapid and interactive visualization over the web. In addition,

improved libraries for dynamic generation of phylogenetic trees are also now available, for example jsPhyloSVG (Smits and Ouverney, 2010), Phylocanvas (http://phylocanvas.org/), EvolView (He et al., 2016), and TnT (which builds on D3 libraries) (Pignatelli, 2016). Thus, with the latest capabilities of web browsers, real-time and interactive visualization of genomic data will be a powerful tool for bioinformatics analyses and our efforts for development of web-based visualization of microbial genomes will be discussed further in Chapter 3.

1.5. Virulence factors and infectious disease pathogenesis

Virulence factors (or VFs) are gene products that allow a pathogen to cause disease in a host by allowing the organism to (i) enter the host, (ii) multiply in the host, (iii) avoid host defense barriers, or (iv) damage the host (Smith, 1977). In order to identify VFs, Stanley Falkow proposed “Molecular Koch’s Postulates” (Falkow, 1988) which state that the virulence factor must be present in all pathogenic strains and absent from non-pathogenic strains, inactivation of the VF must attenuate virulence in an animal model, and re-constituting the VF should re-establish virulence. However, Falkow himself later stated that these postulates have limitations and can be controversial; VFs may only play a role in a specific host for example, and the distinction between pathogen and non-pathogen strains is not always clear (Falkow, 2004). As stated earlier in section 1.3.5, VFs have also been shown to spread horizontally via GIs (Ho Sui et al., 2009), which can lead to the acquisition of virulence in previously designated non-pathogens. Therefore, defining the terms virulence, pathogen, and VF is quite complex.

VFs play an important role in the progression of infectious diseases, but some VFs are present only in pathogenic species, while others are also seen in non-pathogenic species (Zhang et al., 2003). This has led some to further classify VFs as “true virulence genes”, “virulence-associated genes”, or “virulence-lifestyle genes” to distinguish between those VFs that are only ever seen in pathogens and are directly involved in interactions with the host and causing damage, from VFs that are more commonly shared among pathogens and non-pathogens and play a larger role in survival, multiplication, or host immune evasion (Wassenaar and Gaastra, 2001). As mentioned, virulence is also a very contextual phenomenon. Although an organism may have all the necessary VFs to cause

disease, virulence is directly dependent on interactions with the host (Casadevall and Pirofski, 2001), as demonstrated by pathogens that lack the ability to cause disease in immune hosts. Conversely, non-pathogens have also been seen to cause disease in immuno-compromised hosts. Furthermore, the progression of disease may also depend on the environment, for example, a pathogen may favour certain body sites that have ideal conditions for growth. Thus, the virulence of a pathogen is dependent not only on the presence of VFs, but also host-pathogen interactions, and the environment. Consequently, it is extremely important to assess putative new VFs with caution. And though there is no single definitive definition of a VF, we can classify genes based on their function to attain a better understanding of the role of these important genes in the pathogenicity of an organism. Even so, certain VFs have been well-characterized and we have a notable understanding of the underlying mechanisms of their role in disease progression. In the following section, I will describe some of the major types of VFs and their role in causing disease.

1.5.1. Major types of virulence factors

Adhesins facilitate the interaction between pathogen and host. Adhesins bind to specific host cell ligands and are typically responsible for host and tissue specificity of pathogens. For example, fimbrial adhesins (or pili) from various *E. coli* strains have been shown to bind specifically to different receptors on epithelial cells of different hosts (Klemm and Schembri, 2000; Stromberg et al., 1990). **Invasins** are VFs that mediate entry of the pathogen into host cells, either by disrupting the host cytoskeleton or signalling pathways, such as the *ipa* invasion proteins in *Shigella flexneri* (Menard et al., 1996). **Toxins** are secreted factors that directly cause damage to the host. Type I toxins act by stimulating the immune system to induce massive inflammation and can lead to toxic shock and are found in *S. aureus* and *S. pyogenes* (Marrack and Kappler, 1990). Type II toxins are capable of damaging phospholipids or forming pores in cell membranes. *L. monocytogenes* listeriolysin O is capable of disrupting the phagosomefor escape into the cytoplasm of a host cell (Cossart et al., 1989). Type III toxins are intracellular toxins that can alter or regulate the production of critical host factors. For example, *Bordetella pertussis* adenylate cyclase toxin produces cyclic adenosine monophosphate (cAMP), which then accumulates in the host cell and disrupts the proper function of immune cells

and ultimately weakens the immune response (Boyd et al., 2005). Other VFs play a role in **evasion of host defences**, for example, by inactivating antibodies or forming extracellular capsules or biofilms to prevent phagocytosis. **Iron uptake systems** are another important class of VFs as iron is an important nutrient that can be very limited in the environment. For example, siderophores have been discovered that have high affinity for iron and can remove it from host proteins, such as Enterobactin found in many *E. coli* and *S. enterica* serovar Typhimurium strains (Pollack and Neilands, 1970). **Transport systems** (also known as secretion systems) are also a vital class of VFs that help transport other VFs across membranes. There are seven known secretion systems that are sub-divided into two general categories: those that secrete factors extracellularly (Type I, II, V, and VII), and those that inject factors directly into a target cell (Type III, IV, and VI). Type IV secretion systems (T4SS) can also be used for conjugation of genetic material into target cells as mentioned previously. Finally, **VF regulators** represent another important class of VFs. These are important in controlling the expression of genes that are required at different stages of the infection process or in response to environmental stimuli. For example, quorum sensing regulators *las* and *rhl* in *P. aeruginosa* control the production of biofilms once the cell density passes a certain threshold level (Pesci et al., 1997).

1.5.2. Computational resources for the identification of virulence factors

There are multiple online databases that contain annotations of VFs from experimentally verified sources. The most comprehensive database is the Virulence Factor Database (VFDB, http://www.mgc.ac.cn/VFs/) that covers 30 genera of pathogenic bacteria in the latest release (Chen et al., 2016). MVirDB (http://mvirdb.llnl.gov/) is another resource which has curated VF annotations but also integrates annotations from other resources including Tox-Prot, SCORPION, PRINTS, VFDB, TVFac, in addition to annotations of GIs from Islander, resistance genes from ARGO, and virus proteins from VIDA (Zhou et al., 2007).

While the definition of VFs is fairly complex and represents many different unrelated classes of proteins, there is a growing need for rapid identification of VFs in

newly sequenced genomes if they are closely related strains of very well-studied pathogens. Current algorithms for the computational prediction of VFs are summarized in Table 1.3 and typically involve BLAST-based sequence similarity searching. MVirDB and VFDB both have BLAST-based search interfaces to identify homologs against the respective databases of curated VFs. There are other BLAST-based methods that are more specific, such as SIEVE that only predicts type III and IV secreted effector proteins (McDermott et al., 2011) and others only for type III secreted proteins (Samudrala et al., 2009; Yang et al., 2010) by detecting very specific motifs. Such BLAST-based homology searches against databases of known VFs can be combined with other types of sequence-based searches like subcellular localization prediction (e.g. using PSORTb (Yu et al., 2010)), or motif searching for VF-associated motifs to improve accuracy and confidence in predictions. They can also be combined with comparative genomics analyses that identify regions of the genome unique to pathogens versus a collection of non-pathogens that may encode VFs, essentially searching for PAIs. Another method for identification of VFs is VirulenceFinder, which is used specifically for the identification of particularly well-studied VFs in *E. coli*, Enterococcus species, and *S. aureus* from short read sequencing data to rapidly assign subtypes that are defined based on the sequence of the specified VFs (Joensen et al., 2014). VirulentPred is the only algorithm that computationally predicts multiple types of VFs across a broad range of species using a machine learning SVM that is trained on a dataset of known VFs from 12 pathogen genera and has reported accuracy of 81.8% (Garg and Gupta, 2008).

The above described methods primarily predict the presence of well-known VFs, but alternatively WGS data can also be manipulated to predict genes involved in an organisms' toxicity and ability to cause disease, as shown in the example of methicillin-resistant *Staphylococcus aureus* (Laabei et al., 2014). Such an analysis, similar to a genome-wide association study (GWAS), integrates genomic variations such as SNVs and insertions/deletions (indels) with phenotypic expression and can be used to identify VFs (including novel ones) important in pathogenesis, and will become increasingly important for the computational identification of VFs. This study also showed that identified SNVs and indels that affect toxicity can be used to further accurately predict toxicity levels in other strains (Laabei et al., 2014).

In all, the computational prediction of VF is not straightforward and few resources exist for the computational detection of VFs in microbial genomes. Most methods are based on sequence similarity searches that require a certain level of manual interpretation as virulence is dependent on so many different factors. Alternatively, identification of pathogen-associated genes (genes only ever seen in pathogens and never in non-pathogens, described in more detail in section 2.2.3) may aid such analyses to identify genes that may be virulence-associated and could be potential drug targets (Ho Sui et al., 2009).

Table 1.3 Tools for computational identification of VFs

Resource	URL	Details
MVirDB	http://mvirdb.llnl.gov/	BLAST search against all curated VFs
SIEVE	http://www.sysbep.org/sieve	BLAST search to predict type III and IV secreted effectors
VFDB	http://www.mgc.ac.cn/cgi-bin/VFs/jsif/main.cgi	BLAST search against all curated VFs
VirulenceFinder	http://www.genomicepidemiology.org	BLAST search to predict *E. coli*, Enterococcus and *S. aureus* VFs
VirulentPred	http://bioinfo.icgeb.res.in/virulent/	SVM predictor trained on known VFs from 12 pathogen genera

1.6. Antimicrobials and antimicrobial resistance genes

Antimicrobials are agents that either kill or inhibit the growth of microorganisms, including bacteria, viruses, fungi and protozoa, and can be natural, synthetic or semi-synthetic. They are also known to be anciently developed (Wright and Poinar, 2012). Antibiotics are a specific type of antimicrobial agent that kill bacteria and are typically derived from antibiotic-producing bacteria, but can also be synthetic. Particular genera of bacteria, such as Streptomyces, have evolved to produce antibiotics to inhibit or literally destroy competing bacteria in environments with limited nutrients (Aminov, 2009), or to function as cell-to-cell signaling molecules (Aminov, 2009; Linares et al., 2006). Antibiotics have been discovered from such sources and have been used to clinically treat infections in humans and other animals, especially livestock raised for food. Most antibiotics target either (i) cell wall biosynthesis, (ii) protein synthesis, or (iii) DNA replication and repair (Walsh, 2000), but other types of antimicrobial drugs can also target (iv) cell membranes

(Epand and Vogel, 1999) or (v) folate synthesis (Sköld, 2000) or (vi) can intercalate DNA (Armstrong et al., 1970). Table 1.4 lists the major classes of antimicrobial drugs and the pathways that they target to inhibit or kill microbes. Other miscellaneous antimicrobials that are not represented by any of these major classes but may be found in our dataset are fosfomycin, isoniazid, triclosan, and pyrazinamide.

Table 1.4 Major antimicrobial agents and targets

Antimicrobial class	Examples	Target pathway
Lipopeptides	Polymyxin, daptomycin, meonomycin	Cell membrane
Peptides	Bacitracin, gramicidin, defensin, actinomycin	Cell membrane
Glycopeptides	Vancomycin, teicoplanin, bleomycin	Cell wall biosynthesis
β-lactams	Penicillin, cephlasporin, carbapenem	Cell wall biosynthesis
Aminocoumarins	Novobiocin, clorobiocin	DNA replication and repair
Diaminopyrimidines	Trimethoprim, brodimoprim, tetroxoprim	DNA replication and repair
Fluoroquinolones	Ciprofloxacin, nalidixic acid	DNA replication and repair
Sulfonamides	Sulfadiazine, sulfamethoxazole	Folate synthesis
Sulfones	Dapsone	Folate synthesis
Acridine dyes	Acriflavin	Intercalates DNA
Organoarsenics	Arsphenamine	Not known
Aminoglycosides	Gentamicin, streptomycin, kanamycin	Protein biosynthesis
Elfamycins	Pulvomycin, enacyloxin	Protein biosynthesis
Lincosamides	Lincomycin, clindamycin	Protein biosynthesis
Macrolides	Erythromycin, carbomycin	Protein biosynthesis
Oxazolidinones	Linezolid	Protein biosynthesis
Phenicols	Chloramphenicol, florfenicol, azidamfenicol	Protein biosynthesis
Pleuromutilins	Tiamulin, azamulin, pleuromutilin	Protein biosynthesis
Rifamycins	Rifampin, rifaximin, rifabutin	Protein biosynthesis
Streptogramins	Pristinamycin	Protein biosynthesis
Tetracyclines	Tetracycline, doxycycline, glycylcycline	Protein biosynthesis
Nucleosides and aminonucleosides	Tunicamycin, streptothricin, puromycin	Wide ranging: cell wall and protein biosynthesis

1.6.1. Ancient and widespread resistance against antimicrobials

Antimicrobial resistance (AMR) genes are genes that can halt the action of antimicrobial antimicrobial agents in order for the microbes to survive and continue to grow. against virtually every antimicrobial drug emerges after it has been used to clinically treat infections (Clatworthy et al., 2007; Davies and Davies, 2010; Wright and Poinar, 2012) and has the ability to spread rapidly via HGT, but resistance itself is an ancient phenomenon and has not simply evolved since the introduction of clinical antimicrobial drugs (Perry et al., 2016; Wright and Poinar, 2012). It is widespread, found in samples from numerous environments such as soil, sludge, microbiota from different animals Wintersdorff et al., 2016), and even permafrost samples dating back 30,000 years (D'Costa et al., 2011) and caves that have been isolated from over 4 million years et al., 2012) have shown the presence of a wide array of AMR genes. Current environmental soil samples exhibit high levels of resistance against all classes of antibiotics, including extensively used antibiotics and those recently developed, with no differences between natural and synthetic drugs (D'Costa et al., 2006). Many of these AMR genes come from antibiotic-producing organisms, as they must have co-evolved resistance genes to protect themselves, or commensal bacteria. For example, plasmid-encoded quinolone resistance genes that are widely distributed among Enterobacteriaceae species were found to originate from *Shewanella algae* and different *Vibrionaceae* species (Poirel et al., 2005a; Poirel et al., 2005b). These resistance genes can be mobilized on self-replicating plasmids, transposons, integrons, prophages, or genomic islands (discussed in further detail in section 1.3) to spread throughout other populations horizontally (Normark and Normark, 2002). The mobilization of AMR genes also ancient. For example, phylogenetic analysis of OXA β-lactamases shows the mobilization of these genes from chromosomes to plasmids occurred millions of years on at least two independent occasions (Barlow and Hall, 2002).

Table 1.5 highlights some examples of mobile elements that harbour different types of AMR genes.

Table 1.5 Examples of mobile elements carrying AMR genes

Name of element and associated organism(s)	Size	Protection against
Salmonella genomic island 1 (SGI1)	43kb	Ampicillin, chloramphenicol, streptomycin, sulfonamides, and tetracycline (Boyd et al., 2000)
Staphylococcal cassette chromosome *mec* (SCC*mec*)	24kb	Methicillin (Ito et al., 2003)
AbaR1 of *A.staphy baumannii* AYE	86kb	Aminoglycosides, β-lactams, cotrimoxazole, chloramphenicol, tetracycline (Fournier et al., 2006)
Transposon 917 (Tn917) of *Streptococcus faecalis*	5kb	Erythromycin
IncC plasmid R55 of *Klebsiella pneumoniae*	150kb	Chloramphenicol, florfenicol (Cloeckaert et al., 2001)

Nonetheless, since antimicrobial drugs have been mass produced, overused and misused around the globe, there is evidence that the abundance of AMR genes in the environment is increasing (Knapp et al., 2009). More importantly, the clinical use of antimicrobials has especially led to the selection of numerous drug-resistant disease-causing pathogens, which can be observed merely years after the introduction of a new drug (Clatworthy et al., 2007; Davies and Davies, 2010). Notably, multi-drug resistant strains are also being observed in human infections. One of the biggest threats today is carbapenem-resistant Enterobacteriaceae (CRE) strains (including *E. coli* and *Klebsiella pneumoniae*) that have become resistant to virtually every antibiotic that is available today (Gupta et al., 2011). Based on data from the National Nosocomial Infection Surveillance (NNIS) system, CRE strains were exceptionally uncommon before 1992, but have since spread across the United States and around the world (Gupta et al., 2011). There are very limited treatment options against CRE infections and the mortality rate is especially high (Patel et al., 2008).

In general, the prevalence of AMR across the globe is increasing at alarming rates due to the rapid spread of many diverse AMR genes from the environment, commensal and pathogenic species via HGT, and is being out-paced by the discovery of new antimicrobial therapies (Theuretzbacher, 2011).

1.6.2. Mechanisms of resistance

The main mechanisms of antimicrobial resistance are (i) active efflux of the drug, (ii) decreased cell permeability for drug, (iii) enzymatic inactivation or modification of the drug, (iv) inactivation or modification of the drug target, or (v) offering an alternative target that is resistant to the drug (Normark and Normark, 2002; Walsh, 2000). I will further discuss each of these mechanisms below.

Efflux pumps, found commonly across microbial life as well as eukaryotes, are transport proteins that remove toxic compounds from cells, including antimicrobials (Van Bambeke et al., 2000). These pumps are generally part of an operon that encodes all the necessary structural components as well as regulatory genes to control expression, and resistance tends to be mediated by over-expression of the pump (Webber and Piddock, 2003). Thus, mutations in regulatory genes of efflux pumps can often lead to resistance. Efflux pumps can be specific for the transport of one particular molecule, or can be broad-spectrum to remove several unrelated molecules, which are often associated with multi-drug resistance. The presence of “intrinsic resistance” in certain species can largely be attributed to the presence of one or more efflux pumps (Webber and Piddock, 2003). In general, this type of resistance mechanism is likely to be anciently derived as most efflux pumps are fairly conserved and encoded on the chromosome rather than mobile elements (Martinez et al., 2009; Poole, 2005; Webber and Piddock, 2003). Another mechanism of “intrinsic resistance” is through decreased cell permeability for the drug (Normark and Normark, 2002). Mutations or loss of porins can reduce the rate of or prevent drugs from entering the cell (Nikaido, 2001).

A more specialized mechanism of resistance is through the enzymatic inactivation or modification of the drug. A number of enzymatic reactions can accomplish this (Wright, 2005). For one, hydrolysis (cleavage of a molecule by the addition of water) can destroy antibiotics such as β-lactams and macrolides that have susceptible amide and ester groups. More frequently, enzymes covalently transfer particular groups (such as acyls, phosphoryls, thiols, glycosyls, and nucleotidyls) to modify the drug and impair binding to targets. For example, aminoglycosides can be modified by acetyltransferases, phosphotransferases, or nucleotidyltransferases to confer resistance. Although it is not frequent, reduction/oxidation of the drug can also be used to confer resistance, such as

the *tetX* gene that modifies tetracycline drugs (Speer et al., 1991). Finally, lyases can also be used to cleave antimicrobial drugs without hydrolysis or oxidation, such as Vgb, which cleaves streptogramins (Mukhtar et al., 2001). Importantly, these drug modification enzymes can be shared via HGT.

Alternatively, the drug target can be modified instead of the drug, which generally arises through spontaneous mutation of the drug target gene (Lambert, 2005). But this may come at a cost to the organism as these genes typically play essential roles in various cellular processes, so modification of these molecules to resist antimicrobials must be balanced with retaining function. In the case of fluoroquinolone resistance, mutations in DNA gyrase can prevent or lower affinity of binding of the drug and can be found in different subunits of this molecule (Komp Lindgren et al., 2003). Resistance can also be achieved by gaining an alternative drug target molecule that is resistant and can therefore replace the original drug target to perform its intended function in the cell (Lambert, 2005). These genes are very commonly shared via mobile elements. A common example of this is the acquisition of the *mecA* gene, which encodes an alternative penicillin-binding protein 2a that is resistant to β-lactam drugs.

In summary, AMR is a common phenomenon in bacteria through a variety of mechanisms, many of which are able to spread rapidly between unrelated organisms through HGT. Thus, it is important to identify and characterize such acquired genes when using WGS, especially in the framework of infectious disease to assess a pathogen's repertoire of AMR genes, and this dissertation will demonstrate efforts to improve such classification. Not only this, but because no comprehensive study has been previously conducted to determine whether these AMR genes are disproportionately associated with mobile elements over the rest of the chromosome, Chapter 4 of this dissertation will describe an analysis of the trends of AMR genes with mobile elements to improve insight into the spread and evolution of resistance.

1.6.3. Computational resources for the identification of resistance genes

Similar to VFs, AMR genes represent a large collection of diverse proteins and computational methods for prediction of AMR genes are also limited even though there an increasing need for such methods. The main algorithms are summarized in

Table 1.6. The first developed method was ResFinder and it detects AMR genes that can be acquired (i.e. not chromosomal mutations that confer resistance) from short read sequences using a BLAST search against known AMR genes (Zankari et al., 2012). It's important to note that the searches require 98-100% identity over at least 2/5 of the length of the protein so any hits are by default very similar to known AMR genes. The Resistance Gene Identifier (RGI) was developed to further improve BLAST-based AMR gene prediction by also incorporating prediction of chromosomal mutations that confer resistance, while using different predictive models, cutoffs and filters based on the type of AMR gene (McArthur et al., 2013). It can also be run in "strict" or "loose" mode, that can change the cutoffs to be more or less stringent. The accuracy of the RGI has been shown to be very consistent with phenotypic assays of resistance. RGI also provides links to the Antibiotic Resistance Ontology (ARO) for every prediction to aid in understanding the relationships between predicted AMR genes and the drugs they confer resistance against and is a valuable resource for interpretation. ARG-ANNOT is another BLAST-based method that is meant to be used as a standalone tool rather than a web server for more private analyses (Gupta et al., 2014). ARGs-OAP is more recently developed and aims to provide AMR gene predictions from metagenomic samples, also using a BLAST similarity search (Yang et al., 2016).

In summary, AMR gene prediction is complex due to the diversity in genes and mutations that confer resistance, but tools such as RGI that employ different predictive models for the various types of AMR genes are very accurate in detecting close homologs of curated AMR genes in microbial genomes. Computational analysis of AMR gene content of pathogens in particular may become an integral component in infectious disease genomics studies.

Table 1.6 Tools for computational identification of AMR genes

Resource	URL	Details
ARG-ANNOT	http://www.mediterranee-infection.com/article.php?laref=282&titer=arg-annot	Standalone tool for BLAST search against collection of known AMR genes
ARGs-OAP	http://smile.hku.hk/SARGs	Galaxy server for BLAST search of metagenomics data against known AMR genes
ResFinder	http://www.genomicepidemiology.org	BLAST search against database of known acquired AMR genes
RGI	https://card.mcmaster.ca/analyze/rgi	BLAST search against database of known acquired AMR genes and mutations with different models for types of AMR genes

1.7. *Listeria monocytogenes* as a model for studying GIs

L. monocytogenes is a Gram-positive facultative intracellular bacterium found ubiquitously in nature, especially in soil, water, vegetation and can be shed through feces of asymptomatic animal carriers (Lyautey et al., 2007a; Lyautey et al., 2007b; Vivant et al., 2013; Weis and Seeliger, 1975). This pathogen is a serious public health concern as it has been linked to multiple food-borne outbreaks across the globe due to its ability to grow in food preservation conditions such as temperatures as low as 4°C, extreme pH, and high salt concentrations (Cossart, 2011). *L. monocytogenes* infections can lead to listeriosis in immuno-compromised individuals including pregnant women, infants, and the elderly. Listeriosis is characterized by the crossing of the pathogen from the intestinal tract into the lymph nodes via M cells and spreading into the bloodstream, liver and spleen. From there, it can cross the blood-brain barrier to cause meningitis, or the placental barrier in pregnant women into the growing fetus to cause severe illness with high fetal mortality rates (up to 30%) (Swaminathan and Gerner-Smidt, 2007). In the United States, it estimated that *L. monocytogenes* causes more than 1,600 invasive infections and 266 deaths annually and is one of the leading causes of death from food-borne illnesses (Scallan et al., 2011). In Canada, this pathogen has also been the cause of multiple food-borne outbreaks, including a large outbreak in 2008 linked to cold cuts that resulted in 22 deaths (Weatherill, 2009).

1.7.1. Major virulence determinants of *Listeria monocytogenes*

The internalin (*inl*) genes are a family of proteins found on the surface of *L. monocytogenes* that promote adherence and invasion into host cells and all contain a domain with a number of leucine-rich repeats (LRRs). ***InlA*** is a protein that binds to e-cadherin on mammalian cells via LRRs and promotes entry into the host cell (Gaillard et al., 1991). Truncation or loss of *inlA* attenuates virulence of *L. monocytogenes* so this gene is essential for virulence (Nightingale et al., 2008). *InlB* can also mediate entry into host cells, but binds to the hepatocyte growth factor (or Met) (Braun et al., 1998). Upon internalization, the bacterium is located within a phagocytic vacuole. The main VFs for intracellular survival are all located within a cluster of six genes named *Listeria* pathogenicity island 1 (or LIPI-1), all of which are essential genes for virulence (Cossart et al., 1989; Kocks et al., 1992; Raveneau et al., 1992; Smith et al., 1995). ***PrfA*** is the key regulatory protein that activates expression of other VFs in LIPI-1, such as the hemolytic toxin (or **listeriolysin O**), that form pores in the vacuole for escape (Vazquez-Boland et al., 2001). The process of escape also requires the interacting phospholipases ***plcA*** and ***plcB,*** as well as a zinc metalloproteinase ***mpl*** that activates *plcB* (Marquis et al., 1997; Smith et al., 1995). Once the bacterium is in the cytoplasm, it can continue to replicate and spread to adjacent cells. ***ActA*** is a surface protein that plays a role in actin-based motility for cell-to-cell spread by polarized expression to promote asymmetric actin polymerization, which can propel the cell in the opposite direction (Kocks et al., 1993). By this mechanism, *L. monocytogenes* can invade adjacent cells into a secondary phagosome and begin the intracellular lifecycle again.

Other VFs include *inlC*, which is secreted inside infected cells where it plays a role in dampening the innate immune response by preventing activation of NF-κB (Gouin et al., 2010). A regulatory gene, *sigmaB*, is responsible for regulating expression stress-response genes in reaction to environmental stresses such as high salt or acid (Abram et al., 2008). Another regulator, *virR*, regulates genes involved in modification of cell wall and membrane that can confer resistance to antimicrobial peptides (Thedieck et al., 2006). The *Fur* regulon modulates the uptake of ferric iron into the cell (Ledala et al., 2010).

1.7.2. Genomic variation of *Listeria monocytogenes*

The *L. monocytogenes* genome is roughly 2.94 Mb in size. Although the genome is not very large, *L. monocytogenes* discriminates into 13 serovars (or seroypes) based on the type of somatic (O) and flagellar (H) antigens that are also further very diverse. Two different strains from the same serovar can have more than 29-30,000 SNVs (Becavin et al., 2014), while strains from different serovars have been shown to have more than 105,000 SNVs (Nelson et al., 2004). Of note, only four of these serotypes cause most listeriosis cases (1/2a, 1/2b, 1/2c, and 4b); serovar 4b causes most of the maternal/fetal listeriosis, while 1/2a is most commonly found associated with food and food production equipment (Cossart, 2011). Additional typing is commonly performed for *L. monocytogenes*, including the generation of MLST profiles based on sequence analysis of the combination of alleles of seven housekeeping genes (Liu, 2006). Using the genes analyzed for the MLST profile, isolates can be further distinguished into clonal complexes that have allelic differerences in at most one gene from each other (Feil, 2004).

Not only is there significant variation at the nucleotide level, but there is also considerable HGT seen in *L. monocytogenes*, especially from the insertion of phage. Nearly 500 *Listeria*-specific phages have been discovered (Hagens and Loessner, 2015). In fact, phage typing became a common tool to classify strains of *L. monocytogenes* (Loessner, 1991; McLauchlin et al., 1986), but not all strains can be phage typed and there have been reports of limited reproducibility using methods developed by different groups (McLauchlin et al., 1996). In addition, nine integration hotspots have been previously identified for the incorporation of accessory genes (Kuenne et al., 2013). The LIPI-1 pathogenicity island encoding key intracellular survival VFs is thought to have very ancient origins. It does not have any detectable features of GIs (no mobile elements, sequence composition biases, flanking direct repeats, not integrated near a tRNA gene), but is sporadically distributed in the Listeria phylogeny. Notably, LIPI-1 is present in all pathogenic Listeria species (*L. monocytogenes* and *L. ivanovii*), and either absent (*L. innocua* and *L. welshimeri*) or disrupted (*L. seeligeri*) in non-pathogenic species (Chakraborty et al., 2000; Vázquez-Boland et al., 2001). It is not likely that this cluster of genes arose from convergent evolution across multiple Listeria species, so LIPI-1 is probably an anciently acquired GI that lost its mobility. Novel GIs have also been reported,

including the 50 kb Listeria GI 1 (LGI1) that was discovered in Canadian food-borne outbreak isolates and encodes various potential VFs, such as T4SS proteins, and a multidrug efflux pump (Gilmour et al., 2010).

Because of the notable variation between strains of *L. monocytogenes* (including small and large-scale differences), and because it has been associated with many outbreaks, I selected this organism as a model to examine the importance of GI prediction for infectious disease pathogens, including in the context of real outbreaks.

1.8. Goals of present research

When beginning this project, the Brinkman laboratory had published one of the first studies using genomic epidemiology in collaboration with the BCCDC to understand the spread of *Mycobacterium tuberculosis* in a small rural city in British Columbia (Gardy et al., 2011). This study revealed many insights from using WGS in these types of investigations. For one, there was a need for improved methods for characterization of VFs, AMR genes, and mobile elements, like GIs. At the time, no methods existed to identify such elements rapidly in incomplete genomes. To address this, my research focused on improving the characterization of microbial genomes and applying these methods to improve our understanding of infectious disease pathogens in general and in the context of outbreaks.

IslandViewer is web server for GI prediction that was previously developed in the Brinkman lab. In Chapter 2, I describe improvements to IslandViewer that expand the capabilities of the web server to characterize microbial genomes through the integration of curated and predicted homologs of VFs, AMR and pathogen-associated genes (Dhillon et al., 2013). Chapter 3 presents further development of IslandViewer's visualization interface that improves the interactive interpretation of GI predictions in the context of important gene annotations directly through the web server (Dhillon et al., 2015).

By developing these improved methods, I was able to use the rich datasets of GIs and AMR genes over a large collection of diverse genera to perform the first large-scale analysis of the association of AMR with mobile elements. I was also able to further break

down the dataset to study the trends of mobility at the level of AMR classes and mechanisms of action using the ARO. This work is presented in Chapter 4 and provides key initial data for further study in the evolution and spread of AMR, including risk assessment for AMR transmission.

In Chapter 5, I show an example analysis of studying GIs in the context of infectious disease outbreaks by using *L. monocytogenes* as a model to investigate real outbreaks to study the spread of GIs that shows this type of analysis could be useful. I've also applied methods to investigate the genomic determinants of cold tolerance of this pathogen in Chapter 6, with no prior knowledge of the role of SNVs or GIs in this phenotype.

Overall, this work aims to improve the ability to perform genomics-based analyses for the interpretation of genetic changes in microbial species that may impact medically-relevant phenotypes such as virulence and resistance. This is especially important as WGS becomes more and more commonplace and an essential component of infectious disease studies. The presented select analyses provide a basis for the study of AMR and GIs that could have practical implications.

I have carried out all work and analyses presented in this dissertation, however there are cases where I collaborated with others to complete a task and these cases are clearly outlined where appropriate (specifically in Chapter 2, Chapter 3, and Chapter 6).

Chapter 2.

Improving GI characterization: overlay of virulence, antimicrobial resistance, and pathogen-associated gene annotations

Portions of this chapter have been previously published in the article "IslandViewer update: improved genomic island discovery and visualization", co-authored by B.K. Dhillon, T.A. Chiu, M.R. Laird, M.G.I. Langille, and F.S.L. Brinkman in Nucleic Acids Research, 41(W1) © 2013 Dhillon et al; licensee Oxford University Press, and "IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis", co-authored by B.K. Dhillon, M.R. Laird, J.A. Shay, G.L. Winsor, R. Lo, F. Nizam, S.K. Pereira, N. Waglechner, A.G. McArthur, M.G.I. Langille, and F.S.L. Brinkman in Nucleic Acids Research, 43(W1) © 2015 Dhillon et al; licensee Oxford University Press.

I completed all work presented in this chapter with the following exceptions; the RGI analysis presented below was completed in collaboration with Andrew McArthur, Fazmin Nizam, Sheldon Pereira, and Nicholas Waglechner from McMaster University, and the pathogen-associated genes update was performed in collaboration with Geoff Winsor using previously developed code written by Geoff Winsor, Shannan Ho Sui and Amber Fedynak from the Brinkman lab.

2.1. Introduction

As described in section 1.3.5, GIs play an important role in the spread of a variety of different genes between microbes sharing an environment. In the context of infectious disease outbreaks, it is extremely important to not only evaluate small variations like SNVs between isolates to track their spread, but to consider whether or not the acquisition of new genomic material (i.e. GIs) play any role in increasing the transmissibility, toxicity, or resistance of the offending pathogen. Thus, GI prediction is a critical step when using WGS to investigate infectious disease outbreaks. Not only this, but it is increasingly vital

to rapidly assess the presence of newly acquired VFs and AMR genes to classify predicted GIs as PAIs or REIs that play a role in these processes.

However, virulence is a complex term that reflects an interaction between pathogen, host and the environment. Traditionally VFs were identified as genes that contribute to the pathogenicity of an organism, from adherence to a host cell to secretion of a toxin, and may contribute to pathogenesis of disease only under certain conditions, such as in specific hosts. In addition to this, many “classical” virulence factors, such as adhesions, were also discovered to be present in non-pathogenic bacteria (Pallen and Wren, 2007; Snyder and Saunders, 2006; Zhang et al., 2003) and as such it was proposed that VFs should be called more generally “host-interaction factors” (Holden et al., 2004). Because of this reason, there is great interest in identifying pathogen-associated genes, or genes only ever found in pathogen genomes and never in non-pathogen genomes, as these genes may be most critical for progression of disease. Previous studies have shown that these pathogen-associated genes tend to encode more “offensive” virulence functions, such as toxins and secreted effectors as opposed to “common” functions like iron uptake, antiphagocytosis and motility (Ho Sui et al., 2009). This set of pathogen-associated genes could also help identify novel virulence factors that may not have been investigated previously. For these reasons, annotations of VFs, AMR genes and pathogen-associated genes are all important in the context of GI analysis and were incorporated in IslandViewer.

Currently, most GI prediction methods do not evaluate the presence of such genes to characterize GIs. PredictBias (Pundhir et al., 2008), PIPs (Soares et al., 2012) and GIPSy (Soares et al., 2015) are GI prediction methods that include a step to identify VFs through similarity searching against a set of known VFs. PAIDB v2.0 (Yoon et al., 2015) identifies both VFs and AMR genes to characterize GIs as PAIs and REIs. However, these methods have limitations. For one thing, PIPs is limited to only predicting PAIs so users would have to employ another method to predict other types of GIs. In addition, PIPs does not provide any pre-computed results for standard reference genomes. PredictBias and PAIDB v2.0 are also limited in that the GI prediction is not as accurate as IslandViewer. These methods detect bias in sequence composition, but do not have a comparative genomics component like IslandPick that is available in IslandViewer. Furthermore, the

webserver for PAIDB v2.0 has not been reachable. Therefore, IslandViewer is the most accurate and robust GI prediction resource that would benefit from the integration of information about genes that are involved in virulence and resistance to characterize predicted GIs as PAIs or REIs. This chapter focuses on the integration of curated annotations from external datasets as well as the identification of very close homologs of these genes into IslandViewer that would help researchers quickly characterize and evaluate whether GIs potentially play a role in these processes that are important in medically relevant pathogens.

2.2. Methods

The external gene annotations described below were all mapped against 2,782 complete microbial genomes from the NCBI, dated on September 3rd, 2014, as found in MicrobeDB (Langille et al., 2012) version 88. These annotations are all stored within IslandViewer's backend MySQL database as a separate table for ease of integration into the web server's user interface and visualizations.

2.2.1. Curated external gene annotations

The genomes available in IslandViewer have the basic genome annotation as is required by the NCBI for submission. In order to improve the characterization of GIs, curated annotations of VFs and AMR genes were integrated into IslandViewer. 7,319 distinct VF genes across 203 different genomes have been integrated from the Virulence Factor Database (VFDB) (Chen et al., 2012), Victor's Virulence Factors (http://www.phidias.us/victors/), and PATRIC (Wattam et al., 2014) (summarized in Table 2.1). Only a subset of PATRIC VFs were incorporated, simply those with curated links to literature. It is important to note that some hypothetical proteins may be curated as VFs because they are part of larger clusters of genes that encode mutli-subunit VFs, such as secretion systems, but the exact function of that protein may not be determined (i.e. annotated as "hypothetical" for its gene/protein name). Additionally, curated AMR genes have been included from the Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al., 2013), the most extensive AMR gene database at present. CARD includes annotations from the previous Antibiotic Resistance Database (ARDB) (Liu and

Pop, 2009), whichis no longer maintained. These high-quality annotations cover the most well-studied human pathogens, such as *S. enterica, L. monocytogenes*, and *P. aeruginosa*, and will support improved characterization of predicted GIs as PAIs or REIs directly within IslandViewer.

2.2.2. Expanding coverage of curated datasets

Through the integration of additional gene annotations into IslandViewer, I was able to provide rich annotations of curated VFs and AMR genes for 527 different genomes, but this still represents a very small proportion of the total available genomes in IslandViewer. Typically, curated annotations of VFs and/or AMR genes come from one, sometimes two, very well studied strains of a particular species. And although these same classical VFs and/or AMR genes may be present in closely related strains of the same species, they are just not curated as such. To address this lack of curation in very closely related strains of well-curated genomes, I developed a stringent annotation transfer protocol for VFs. Furthermore, I used the Resistance Gene Identifier (RGI) for identifying close homologs of AMR genes.

Annotation transfer of VFs

For the identification of closely related VF homologs, first, criteria were set to determine which genomes were “closely related” to well-curated genomes to perform the annotation transfer. For one thing, I decided a candidate genome must be of the same genus and species designation as the curated “reference” genome. For *S. enterica* and *Eschericia coli*, I further required the genome to be from the same serovar as the “reference”. Second, I used CVTree (Xu and Hao, 2009) as a distance measure to filter out any genomes within the species designation that were unusually diverged. CVTree uses a composition vector to measure phylogenetic distance based on the whole genome rather than a particular gene and is already incorporated into the IslandViewer pipeline for use in the IslandPick GI prediction method. Thus, all genomes in IslandViewer have pre-computed CVTree distances against all other genomes in the database. Only those genomes with a CVTree distance less than 0.3 to a “reference” genome with curated VFs (and within the same species or serovar) were selected as candidates for a VF annotation transfer.

As an example, Figure 2.1 depicts a phylogenetic tree based on a single gene, DNA gyrase subunit B (or *gyrB*) overlaid with the cumulative CVTree distances calculated from the selected reference, *P. aeruginosa* PAO1. First, the CVTree distances largely follow the phylogenetic distances. Second, all *P. aeruginosa* strains are within a CVTree distance of 0.2, while other Pseudomonas species have distances greater than 0.4. Since the distance cutoff between strains of the same species and strains of different species may differ for other genera, cut-offs of 0.2, 0.3, and 0.4 were tested. Overall, 0.2 was very stringent and removed too many isolates from within a species designation that should be considered close relatives, while 0.4 permitted more distant isolates that were closer to other species designations. Thus, a CVTree distance cutoff of greater than 0.3 was used to filter out any strains within the same species designation that are extremely diverged.

Upon selection of candidate genomes, protein BLAST (Camacho et al., 2009) was used to calculate reciprocal best BLAST hits (RBBHs) for each curated VF between the two genomes with very stringent criteria. Previous studies have shown that for single-domain proteins, the precise function of a protein is conserved down to roughly 40% sequence identity between domains of the same fold, while broad functional classes are conserved with as low as 25% sequence identity (Wilson et al., 2000). For multi-domain proteins, if the combination of domains is the same, the probability of sharing function is high (80%) and increases to more than 90% if the domains are shared across the full length of the protein (Hegyi and Gerstein, 2001). In this case, I required that the RBBHs must share 90% identity over at least 80% of the length of the protein and have an expectation value (e-value) less than 10^{-10}. The requirement for similarity over at least 80% of the length of the protein ensures matches are not between single domains but rather the full protein, and also removes false predictions of truncated VFs that are not likely performing an equivalent function. These extremely stringent cutoffs are important in order to maximize precision/specificity at the expense of recall/sensitivity to ensure the identified VF homologs are very likely true orthologs performing the same function, at the expense of missing some. This widely accepted RBBH criteria tends to identify orthologs, but I err on the side of caution and refer to them as homologs.

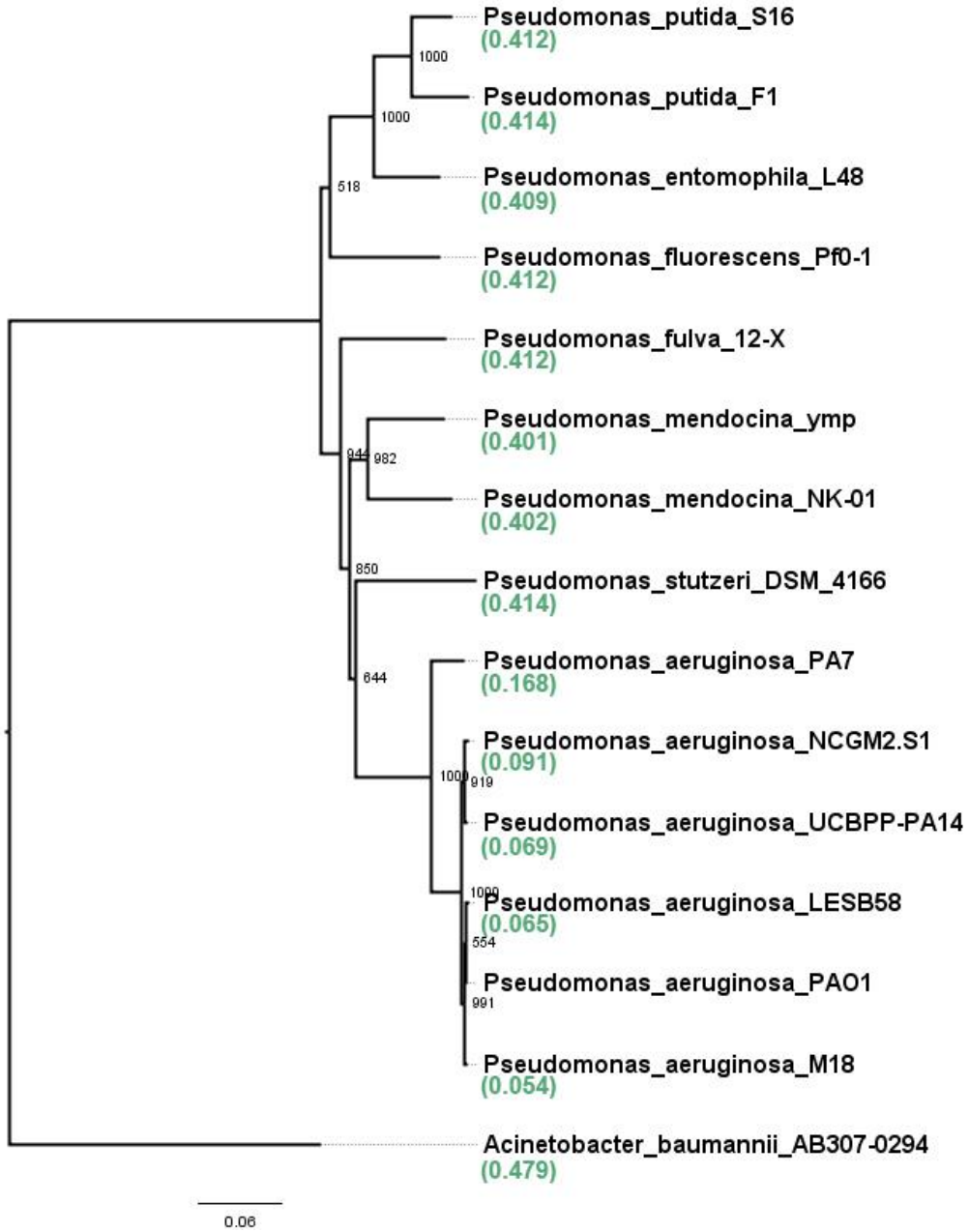

Figure 2.1 Phylogenetic tree of Pseudomonas strains with CVTree distances
Phylogeny based on gyrB sequences of selected Pseudomonas isolates and CVTree distances shown in green as calculated within IslandViewer pipeline.

Evaluation of annotation transfer

Evaluation of this annotation transfer for VFs is not straightforward since the curated VFs generally come from one well-studied genome. In order to evaluate the method, a set of these curated VFs would have to be removed and then tested for how

well they are predicted, but because there are no other genomes closely related enough that have a similar curated VF, they will not be found. This approach did not yield any meaningful evaluation of the accuracy of the annotation transfer. I further attempted to compare against datasets of Signature-Tagged Mutagenesis (STM) libraries. STM is a method in which transposons carrying unique tags are inserted randomly into different genes to create a library of mutants that are tested for growth *in-vivo*. Those mutants that do not survive in the host are thought to be missing key virulence genes, and so STM studies have been used in the past to discover novel virulence genes (Hensel et al., 1995). However, this type of experiment has biases in that it can identify false positive VFs that are generally involved in growth of the organism or have downstream effects on other genes that are involved in virulence, rather than identifying classical VFs, and thus did not provide a solid framework for evaluation. Nonetheless, in light of the fact that the annotation transfer is only performed between very closely related strains of the same species that very likely contain the same VFs, and because I used very stringent filtering criteria for calling VF homologs, the VFs identified through this method were considered to be of high-quality and included in IslandViewer for interpretation by the user.

RGI for AMR homolog detection

All 8.7 million protein sequences from every genome in IslandViewer were run through the RGI in collaboration with the authors of the CARD database (Andrew McArthur's group) since, at the time of this study, it was only available as a web service. I collaborated with the authors to submit our dataset in batch mode using the back end command-line system. RGI was run in "strict" mode so as to be very conservative in identification of homologs of AMR genes. The RGI is based on a protein BLAST (Camacho et al., 2009) search against the CARD database of curated AMR genes with a default e-value cutoff of 10^{-30}, and custom cutoffs for particular classes of AMR genes. For example, certain classes of AMR that are highly conserved would require much higher e-value cutoffs than others. The RGI is also capable of identifying SNVs that confer resistance using HMMer (Eddy, 2011). I parsed all the compiled results for storage in the IslandViewer database and these homologs of resistance genes are presented with different colours and/or labels than curated AMR genes for clarity on IslandViewer visualizations and downloadable files.

2.2.3. Pathogen-associated genes update

Based on a previous analysis to identify pathogen-associated genes, an update was calculated based on a more recent collection of genomes, since pathogen-associated genes can change depending on the genomes that are used in the calculation. To calculate pathogen-associated genes, first every genome must be curated as either being a pathogen or non-pathogen. The original analysis had curated 298 pathogens and 333 non-pathogens using TIGRs Microbial Genome Properties table (Haft et al., 2005) with some manual curation for completeness. However, this resource is no longer supported and the pathogen-status of each newly sequenced genome was not available, even within the genome submission information from the NCBI. Thus, through manual curation of each new genome, I identified 1,292 pathogen genomes and 1,490 non-pathogen genomes to perform an updated pathogen-associated genes analysis. An all-against-all BLAST on the deduced proteomes of all genomes using an e-value cutoff of 10^{-7} to exclude distant homologs. Other cutoffs were tested previously (including 10^{-12} and 10^{-5}) and the trends with pathogen-associated genes were still observed using these other cutoffs. Any genes found to be in three or more different pathogen genera with no detectable homologs in non-pathogens were included in the final set of pathogen-associated genes.

2.3. Results and Discussion

2.3.1. Summary of incorporated annotation datasets

Table 2.1 summarizes the total number of genes with external annotations available in IslandViewer, broken down by the dataset type, and the number of genomes, species and genera that have such information. This table is broken down by the curated or calculated datasets, to show how much more information is available even by using the strictest homolog detection methods. The incorporation of such rich datasets will greatly improve the ability to characterize and interpret GI predictions in IslandViewer.

The three VF resources identified more than 7,000 unique curated VFs from 203 genomes. The data from the three resources overlap in some of the genes that are annotated as VFs, but the number of unique annotations in each database is much greater

(see Figure 2.2), indicating that each resource tends to focus on complementary niches of virulence or organisms of interest and should all be included.

Table 2.1 Summary of annotations available in IslandViewer

Dataset	Number of unique annotations	Number of genomes	Number of species	Number of genera
Curated VFs	7,319	203	62	36
Annotation transfer VF homologs	39,441	483	45	31
Curated AMR genes	2,451	387	90	53
RGI AMR gene homologs	26,460	2,422	1,212	586
Pathogen-associated genes	18,919	920	333	139

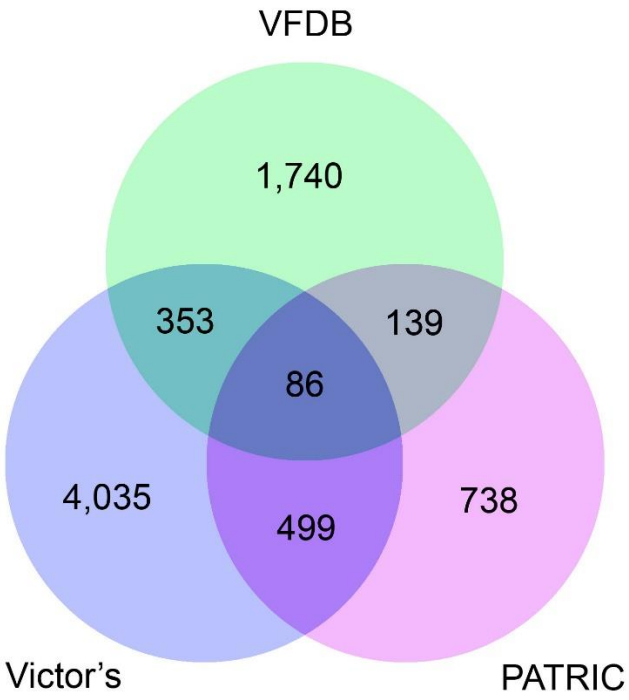

Figure 2.2 Venn diagram highlighting overlap between annotations of curated VFs from VFDB, Victor's and PATRIC

2.3.2. Strict homolog detection greatly increases genomes with available annotations without compromising data integrity

The VF annotation transfer approach allowed the annotation of an additional 39,441 VF homologs in 483 genomes, which more than doubles the number of genomes with VF annotations in IslandViewer. Table 2.2 summarizes for each genus the number of VF homologs detected and in how many genomes, in comparison to the curated dataset. It is evident that this protocol greatly increased the number of genomes with VF annotations and will allow for better interpretation of PAIs in the context of GI prediction of a larger number of genomes. Although a proper evaluation of the accuracy of the annotation transfer was not possible, it is still reliable because of the very stringent criteria that were used for (i) selecting candidate genomes for transferring annotations, and (ii) filtering RBBHs for the most robust homologs. This fully automated protocol has also been incorporated into the IslandViewer update pipeline that downloads newly sequenced genomes from the NCBI database so that VF homologs can be detected in parallel to GI prediction. For clarity, the VF homologs are coloured differently on any visualizations within IslandViewer, are denoted as homologs in any downloadable files, and the website further cautions users to interpret these VF homolog predictions carefully and use them as a guide for generating hypotheses that need further testing.

Table 2.2 Number of VF annotations per genus where annotation transfer increased the number of genomes with annotations

Genus	Number of curated VFs	Number of genomes with curated VFs	Number of VF homologs	Number of genomes with VF homologs
Actinobacillus	115	3	14	3
Bacillus	32	5	19	5
Bartonella	65	4	3	1
Bordetella	106	3	144	2
Borrelia	11	2	7	4
Brucella	450	6	1399	18
Burkholderia	212	12	1111	21
Campylobacter	71	5	111	9
Chlamydia	33	1	1714	55
Chlamydophila	3	3	9	5
Clostridium	25	7	25	7
Corynebacterium	13	1	44	11

Genus	Number of curated VFs	Number of genomes with curated VFs	Number of VF homologs	Number of genomes with VF homologs
Edwardsiella	1	1	2	2
Enterococcus	56	1	30	3
Escherichia	802	26	428	7
Francisella	302	3	571	6
Haemophilus	67	2	182	10
Helicobacter	145	5	4031	51
Legionella	179	5	787	7
Listeria	263	7	3535	36
Mycobacterium	872	12	8311	31
Mycoplasma	11	2	32	14
Neisseria	173	4	894	14
Pseudomonas	293	3	2927	13
Rickettsia	1	1	9	9
Salmonella	675	9	3493	20
Shigella	211	7	147	5
Staphylococcus	114	9	2649	37
Streptococcus	638	29	4876	48
Vibrio	326	6	617	14
Yersinia	206	12	399	15

Figure 2.3 depicts a visual example of results from the VF annotation transfer between two *S. enterica* subsp. enterica serovar Typhimurium strains, a well-curated LT2 strain, and a more recently sequenced 798 strain. The 798 strain had no curated VFs prior to the annotation transfer, and by using LT2's 321 curated VFs, the annotation transfer protocol was able to identify 270 high-quality VF homologs in strain 798. Furthermore, the locations of VF homologs as visualized in Figure 2.3 are very similar to the curated dataset, supporting that these are well-conserved VFs. In all, with this expanded VF homolog dataset, users can now explore the presence/absence of VFs on GIs, such as PAIs, in a much larger collection of genomes. Needless to say, these annotations should still be considered as an initial guide for more in depth analysis due to the highly contextual nature of virulence.

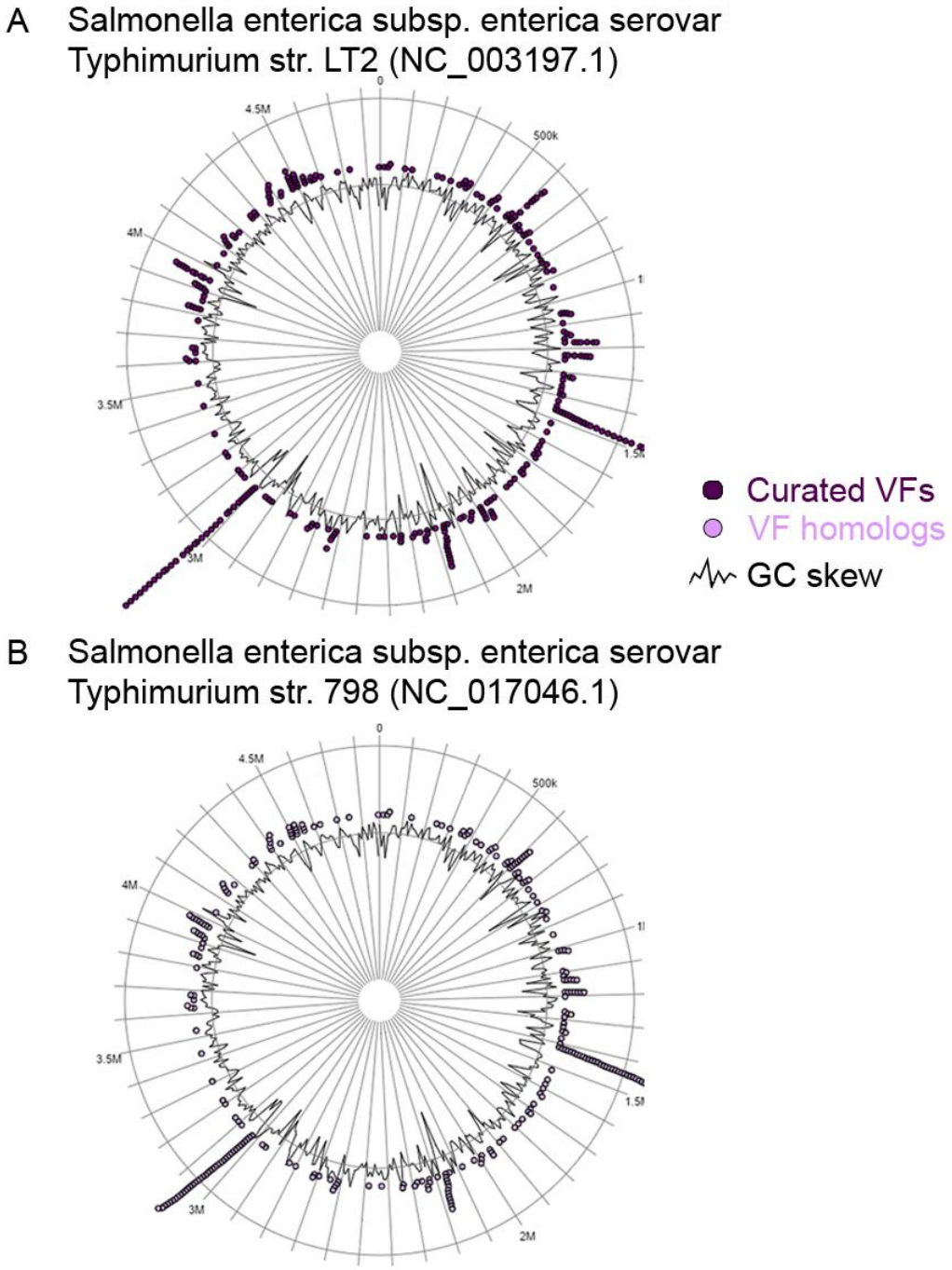

Figure 2.3 Location of VF homologs in comparison to curated VFs between two strains of *S. enterica* subsp. enterica serovar Typhimurium

Circular plots represent the genome and the overlaid scatter plots depict locations of curated (purple) or calculated (light purple) VFs. The GC content skew is also shown as the black line plot. Visualizations were downloaded for each genome from IslandViewer, displaying only the VF tracks. Panel A shows the location of 321 curated VF genes for the LT2 reference strain and panel B shows 270 high-quality VF homologs annotated after the annotation transfer for the 798 strain.

For AMR gene annotation, the RGI was able to identify an additional 26,460 very close homologs of the curated AMR genes in strict mode. Almost every genome analyzed had a predicted AMR gene homolog, in pathogens and non-pathogens alike, illustrating how expansive AMR is within microbial life. Importantly, this large dataset now enables the study of the association of AMR genes with GIs across a large collection of microbial genomes, as was done previously with VFs (Ho Sui et al., 2009), and is presented in detail in Chapter 4.

2.3.3. Updated pathogen-associated genes analysis further improves gene annotations

The updated pathogen-associated genes analysis detected 18,919 genes that are seen in at least three different genera of pathogens and never in non-pathogens, representing a collection of genes that warrant further investigation for their role in the pathogenesis of disease. A majority of pathogen-associated genes (64%) are hypothetical proteins that have likely never been studied for their role in pathogenesis and could represent novel VFs. These genes are now annotated in IslandViewer, visualized in a similar manner as the VFs and AMR genes, and can be interpreted in the context of GI analysis. Pathogen-associated genes found on GIs may be of particular interest to researchers especially in cases where a newly acquired GI may cause a non-pathogen with classical virulence factors to suddenly become pathogenic.

The results from the updated calculation can be found linked from the original web supplement at the following website: http://pathogenomics.sfu.ca/pathogen-associated/. This includes a complete list of pathogen-associated, common, and non-pathogen-associated genes for every analyzed genome for both the original analysis, which included 631 genomes, and the updated analysis of over 2,782 genomes. The classification of genes as being pathogen-associated is now more accurate than the previous analysis simply because so many more pathogen and non-pathogen genomes were included, and will continue to improve further as more genome sequences become available and are included in such an analysis.

2.4. Conclusions

From the perspective of infectious disease outbreaks, while WGS can be used to verify transmission events between individuals and track the progression of an outbreak, this advanced technology also allows researchers to investigate large changes in the genome that may have played a role in initiating the outbreak or may affect the pathogen's response to antimicrobial treatment. GIs are important vectors for the spread of such large clusters of genes between unrelated microbes that may play a role in virulence or resistance, thus identification of such GIs is a critical step in whole genome analyses of outbreak isolates as well as other pathogens in general. Therefore, annotations of VFs, AMR genes and pathogen-associated genes have been integrated into IslandViewer for evaluation alongside GI predictions.

Classical VFs have been annotated using three manually curated databases (VFDB, Victor's and PATRIC), while a conservative VF annotation transfer approach has proven to be beneficial in annotating highly conserved VF homologs in very closely related strains of well-curated genomes. Although the annotation transfer is still limited in the breadth of genomes that can be annotated, it is a notable improvement to build upon the curated datasets. In the future, custom genome submissions can be analyzed using this approach for the identification of VF homologs in parallel with GI prediction. To further improve the annotation of virulence-related genes in IslandViewer, pathogen-associated genes have also been integrated after re-computing based on a larger collection of genomes. These pathogen-associated genes are important because they are never seen in non-pathogen genomes and represent a set of genes that may play a more "offensive" role in pathogenesis of disease and could represent novel VFs that have never been studied for their role in virulence. Overall, IslandViewer provides an expansive collection of VF homologs for interpretation of GI predictions in a large set of pre-computed genomes to identify GIs, or specifically PAIs, that may be of interest.

In addition to this, the incorporation of curated data from the CARD database and predictions using the RGI tool have provided rich annotation of AMR genes in all IslandViewer genomes for the rapid classification of REIs. Of course, as with all computationally predicted annotations, the user must carefully assess the AMR homologs

based on their needs. This dataset was used to analyze trends in the association of AMR genes with GIs, similar to a previous study on the association of VFs with GIs (Ho Sui et al., 2009), and is presented in Chapter 4. The RGI can also be integrated into the existing IslandViewer analysis framework in the future for running on new genomes from the NCBI and custom submissions since a command-line version has just recently been released.

In all, this chapter summarizes the incorporation of external gene annotation data into IslandViewer that is important in the interpretation and classification of GIs, especially from a medical standpoint. Either previously developed resources have limitations in GI prediction capabilities, availability of software, or integration of rich genome annotations and the improvements outlined in this chapter will allow IslandViewer to fill these gaps to be the most useful web server for GI prediction and interpretation.

Chapter 3.

Improved visualization of genomic islands

Portions of this chapter have been previously published in the article "IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis", co-authored by B.K. Dhillon, M.R. Laird, J.A. Shay, G.L. Winsor, R. Lo, F. Nizam, S.K. Pereira, N. Waglechner, A.G. McArthur, M.G.I. Langille, and F.S.L. Brinkman in Nucleic Acids Research, 43(W1) © 2015 Dhillon et al; licensee Oxford University Press.

The development of GenomeD3Plot was led by myself and Matthew Laird, a software developer/programmer in the Brinkman lab who played a vital role in coding the visualization library and re-designing the IslandViewer back-end system while Julie Shay played an important role in the incomplete genomes analysis pipeline.

3.1. Introduction

After improving the characterization of GIs through the integration of additional annotations of VFs, AMR genes and pathogen-associated genes (Chapter 2), I saw the need for building more dynamic and interactive visualizations within IslandViewer, especially with the advancement of web browser capabilities since the introduction of HTML5 (as described in section 1.4.4). Previously, IslandViewer visualizations were pre-computed for every genome in IslandViewer and were not conducive of simple interpretation of GIs in the context of the genome annotations. Web browsers for whole genome visualization do exist, but have limitations. For one thing, many of the visualization tools were developed under the framework of eukaryotic genomes, such as the UCSC genome browser (Kent et al., 2002) and Ensembl genome browser (Fernández and Birney, 2010). There are others, like the Generic Genome Browser (GBrowse) (Stein et al., 2002) and JBrowse (Skinner et al., 2009), that can be applied to the visualization of microbial genomes, but are simply not lightweight for integration into the existing IslandViewer web framework and do not exploit the powerful capabilities of modern

browsers. Such tools are better suited for standalone use and are limited in interactive functionality.

In this chapter, I present the new IslandViewer visualization tool, GenomeD3Plot, developed in collaboration with Matthew Laird, which significantly improves interactive navigation through GI predictions of microbial genomes in the context of rich genome annotations within the IslandViewer framework. As described below, GenomeD3Plot minimizes storage requirements to allow IslandViewer to handle the growing number of new genomes that are being sequenced. Although this chapter focuses on the new visualizations of IslandViewer, the backend of IslandViewer was also completely re-written and re-structured using the Django framework to build a more efficient system that is not only able to handle the improved visualizations, but is able to handle more genomes, especially custom genomes. In addition, a larger cluster of newer compute nodes has been allocated for IslandViewer and a library to schedule and track submissions (MetaScheduler) was developed by Matthew Laird to improve the throughput of custom genomes through the pipeline and improve error handling. Further changes also accommodate the new structure of files within the NCBI microbial genomes server. Overall, these updates to IslandViewer have significantly improved this web server to use the most innovative technology to stay in pace for managing the oncoming growth of complete reference genomes and custom submissions from researchers and provide interactive and dynamic visualizations.

3.2. GenomeD3Plot implementation and features

GenomeD3Plot (Laird et al., 2015) was developed as a lightweight visualization library based on the D3 javascript library (http://www.d3js.org). GenomeD3Plot is able to communicate with the backend MySQL database of IslandViewer to collect genome information and annotations to generate images dynamically when pages are loaded. Previous versions of IslandViewer stored pre-computed Circos plots (Krzywinski et al., 2009) for every permutation of available tracks for each genome and this became a non-trivial storage requirement as I incorporated tracks for VF, AMR and pathogen-associated gene annotations on top of GI predictions from the three methods. For example, a previous version of IslandViewer with roughly 1,800 genomes required 7.3 Gigabytes of space just

to store all the necessary genome images, and this value would be even higher if it included custom genomes. This space requirement is not sustainable as the number of reference genomes available in IslandViewer increases and the demand for custom submissions continually increases. In addition, the static images are also not the most effective means of exploring GI predictions across a genome. These are some of the most important reasons for developing GenomeD3Plot – so that it is scalable in handling an increasing number of microbial genomes by reducing the file storage requirements and providing an interactive interface for comprehensive genome analysis in the context of GI predictions.

An example of the new IslandViewer results page using GenomeD3Plot is shown in Figure 3.1 for *S. enterica* subsp. enterica serovar Typhi str. Ty2 to highlight some of the key features of this tool. GenomeD3Plot visualizations integrate GI prediction results from SIGI-HMM (in orange), IslandPath-DIMOB (in blue) and IslandPick (in green) and are shown in separate tracks, and the outer-most track (in red) is the union of results from all three methods and are considered the IslandViewer GI predictions. The annotations of VFs, AMR genes and pathogen-associated genes are visualized as a scatter plot on top of the GI predictions for improved evaluation of the genome in the context of these important genes. These are coloured dark purple and dark pink for curated VFs and AMR genes respectively, and a lighter shade of purple and pink for any predicted homologs of VFs and AMR genes. Pathogen-associated genes are coloured in orange. The GC content is also shown as the black line plot around the genome for users to visualize any large skews that may be associated with GIs. Users can use the legend to specify which tracks to display, and the tracks dynamically turn on or off on demand without reloading the page (as was the case previously).

To facilitate navigation of all this information across the genome, three separate views are available: circular, vertical, and horizontal views (see panels A, B, and C in Figure 3.1). The circular view is a representation of the complete genome and is the main figure used for selecting regions and orienting the other views. Users can navigate through the genome by clicking on the circular visualization or using the mouse wheel to zoom in/out of regions. Any movements in the circular view dynamically update both the horizontal and vertical views. The horizontal view represents a detailed linear

representation of the location of genes and the vertical view provides the rich annotations of the selection regions. All GI predictions and annotations of VFs, AMR genes and pathogen-associated genes are shown in all three views. Importantly, the entire genome and annotations are loaded from the MySQL database on the first load of the webpage so any navigation around the genome is instantaneous and is a significant improvement from previous versions of IslandViewer. All views can be dynamically resized for different sized screens and are also available for download in high-quality PNG or SVG formats for publications. Because of the dynamic and interactive nature of GenomeD3Plot, we were also able to implement a search feature for users to search for a gene of interest, which is then highlighted/centred. In Figure 3.1, a search for the known VF gene *vexE* has resulted in the highlighting of this gene in all three views. In this case, it is also clear that this gene overlaps with GI predictions from IslandPick (in green) and IslandPath-DIMOB (in blue) and could be a PAI. This could indicate to the user that other genes on this GI could also be of interest and may play a role in virulence and should be further investigated. In this way, GenomeD3Plot is able to enhance a user's ability to efficiently identify important features of GIs directly through the visualizations on the IslandViewer web server. Additionally, users can save a URL that will link back to the exact highlighted positions, acting similar to a "save view" function. Altogether, the GenomeD3Plot visualizations greatly improve the ability to interactively navigate through microbial genomes and study GI predictions in the context of additional genome annotations, including VFs, AMR genes, and pathogen-associated genes that may be of interest, especially in the study of pathogen genomes.

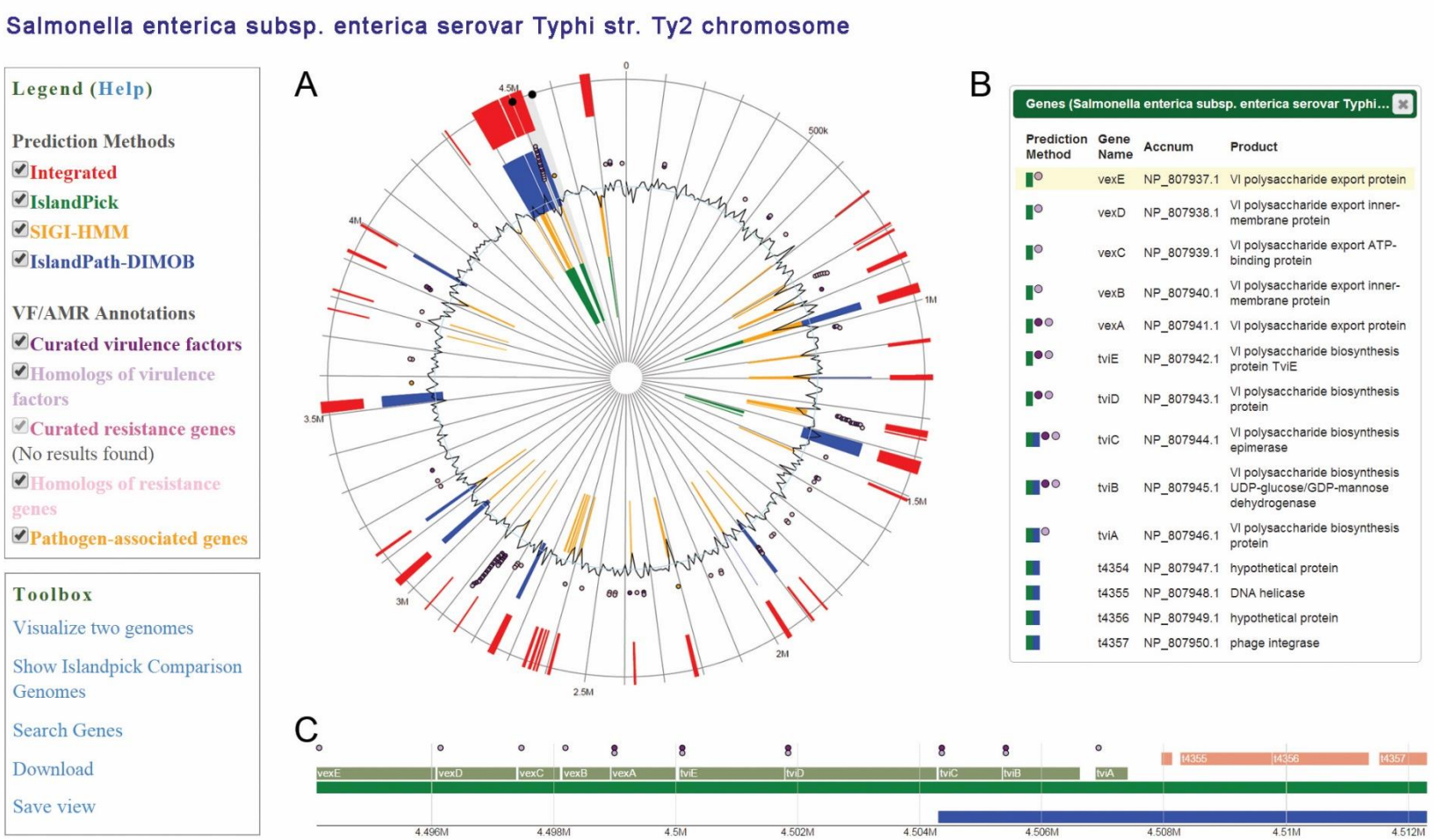

Figure 3.1 IslandViewer 3 results page with GenomeD3Plot circular (A), vertical (B), and horizontal (C) views.
The legend describes the colours of each GI prediction, broken down by method, and the various annotations, and can be used to turn tracks on/off. Circular plot (A) can be used to zoom in/out and select regions of the genome that will be focused in plots (B) and (C) for more details. The toolbox highlights other available features such as search, download, and save views.

A side-by-side comparison page for two genomes was also developed to better facilitate comparative analyses for GI predictions. This feature was highly requested by users, especially for cases where they would like to compare the results for a custom genome against a pre-computed reference genome, or two custom genomes. This comparison page is, however, very limited and it will be important to develop functions for performing comparative analyses in future versions of IslandViewer to handle comparison of tens or even hundreds of genomes simultaneously.

GenomeD3Plot also supports incomplete or draft genomes and clearly denotes contig boundaries in the circular and horizontal views (see example in Figure 3.2). The analysis of incomplete genomes was the most highly requested feature by IslandViewer users. To analyze these genomes, first contigs are ordered against a user-selected reference genome using the Mauve contig mover (Rissman et al., 2009) to build a single concatenated sequence that can be run through the existing IslandViewer pipeline. Any unaligned contigs are simply concatenated to the end of this pseudo-chromosome sequence and are also denoted in the visualizations. Julie Shay played a key role in developing and evaluating this pipeline to determine issues with GI prediction in draft genomes and has shown that many GIs are found at contig boundaries and may be missed in prediction (Shay, 2016). Through a gene function analysis using clusters of orthologous group superfamily designations, Julie also showed that transposons and genes involved in replication, recombination and repair tend to be missing from incomplete draft genomes (Shay, 2016). In addition to this, because the genome is concatenated into a pseudo-chromosome that may be missing sequences in gap regions, any GI predictions that span across these contig boundaries need to be carefully evaluated. But overall, GI prediction in incomplete genomes is still useful, although there may be some missing predictions, and GenomeD3Plot visualizations can help users easily evaluate GI predictions that may be near contig boundaries to determine if they are truly clusters of genes of horizontal origin.

All of these features together are important in allowing IslandViewer visualizations to be used to navigate through genomes at various levels of detail, including virulence and resistance potential, but can also be applied to generally visualize, explore and study microbial genomes. As such, GenomeD3Plot was also published as a library separate

from IslandViewer that can be applied generally for other microbial genome visualization websites and easily customized to display virtually any type of additional data on top of a genome (Laird et al., 2015). This library is licensed under the GNU General Public License (GPL) v3 and is available for download at https://github.com/brinkmanlab/GenomeD3Plot/.

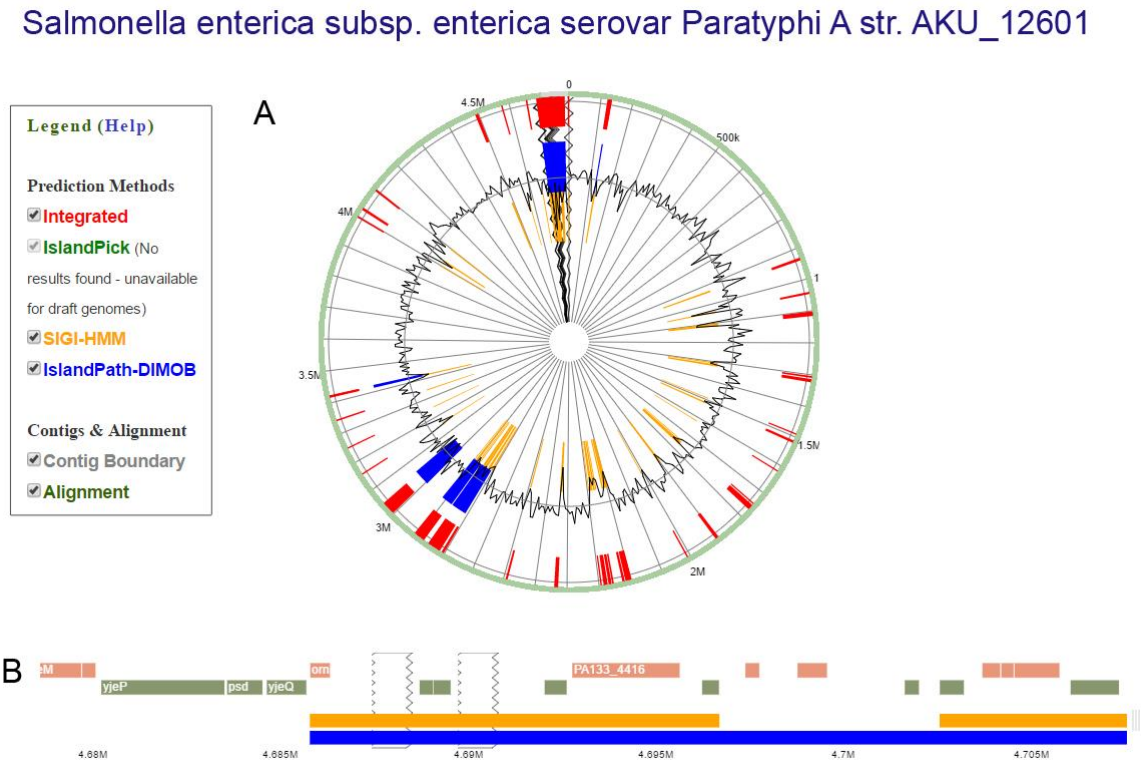

Figure 3.2 IslandViewer results of custom incomplete genome analysis
Incomplete genomes have a customized view using GenomeD3Plot that visualizes the contig boundaries in the ciruclar (A) and horizontal (B) views for the user to interpret GI predictions across these regions. The circular view (A) also provides an extra track along the outside of the plot to show the regions that aligned against the selected reference. The legend is also customized to turn these extra tracks on/off.

3.3. Conclusions

IslandViewer is an important tool that has been widely used by the scientific community to discover GIs in microbial genomes; however, previous versions have been limited in their visualization capabilities. With the addition of a rich collection of genome annotations, visualization of GIs in the context of such data is increasingly invaluable for

interpretation of results but not scalable using previous techniques. GenomeD3Plot was developed to address the limitations of other genome visualization tools in handling IslandViewer results. This tool overcomes the problems of scalability with Circos plots that need to be pre-computed and utilize a significant amount of storage space, given the increasing number of genomes in IslandViewer. Furthermore, it provides an interactive interface that is fast and intuitive in incorporating GI predictions with rich genome annotations, including VFs, AMR genes and pathogen-associated genes to improve the ability of users to easily identify and interpret GIs of interest all within a single web interface. It is also able to handle visualization of incomplete genomes for careful examination of GI predictions and contig boundaries. Another important feature of GenomeD3Plot is that it is available as a separate library that can be integrated into other web tools that require lightweight alternatives for genome visualization libraries. In IslandViewer, GenomeD3Plot can be used to display two genomes side-by-side for comparison of custom genomes against a reference, but this feature is still quite limited and future development efforts will need to focus on the ability to expand this tool to highlight differences in GI predictions over tens, hundreds, or even thousands of genomes in efficient ways. Overall, with the development of GenomeD3Plot, IslandViewer users can now interactively display relevant tracks dynamically, search for genes of interest, easily navigate through prediction tracks in context of a rich set of genome annotations, save views, download publication quality images, and compare results for two genomes. All of these improvements in the visualization of IslandViewer results has greatly improved the ability of users to interpret and interrogate the role of HGT within microbial genomes.

Chapter 4.

Mobility trends of antimicrobial resistance

4.1. Introduction

In Chapter 2, I described the integration of external gene annotations of VFs, AMR genes, and pathogen-associated genes into IslandViewer. VFs have been previously shown to be significantly disproportionately associated with GIs (Ho Sui et al., 2009), so naturally I wanted to combine the rich AMR data and GI predictions to understand the association of AMR genes with these mobile regions of the genome. AMR is known to be very commonly spread using mobile elements such as plasmids, integrons, and transposons; however, no large-scale study has ever shown whether or not AMR genes are strongly associated with such mobile elements. The analysis described in this chapter represents the first comprehensive study of mobility trends of a large collection of AMR genes across phylogenetically diverse genera.

4.2. Methods

A collection of 2,782 genomes was used in this study from version 88 of MicrobeDB (Langille et al., 2012), which includes all complete microbial genomes available from the NCBI on September 3rd, 2014. The trends were also tested on a subset of 164 genomes with a minimum evolutionary distance (substitutions/site) of 0.05 determined based on a previous study by (Ciccarelli et al., 2006) in order to reduce redundancy and remove potential sampling bias from this dataset. The full list of species included in this less-biased collection of genomes is available in Appendix A.

4.2.1. Antimicrobial resistance gene prediction

All protein sequences from each genome were analyzed through the Resistance Gene Identifier (RGI) in strict mode that has a default e-value cutoff of $1e^{-30}$. The RGI identified a set of curated AMR genes (100% match to a curated resistance gene) as well

as potential homologs based on specialized cutoffs for different types of resistance genes. The RGI also screens for SNVs that confer resistance using a HMMer search against a set of hidden Markov Models (HMMs) that are developed by curators and include alignment reference sequences. The discovery mode of the RGI was not used since I did not perform any confirmatory tests of AMR gene predictions and wanted to be confident that any predictions made were very likely to be true AMR genes. The RGI also integrates an ontology, the Antibiotic Resistance Ontology (ARO) and for each AMR gene, the associated ARO term(s) were also extracted for investigation. By using the ARO, it was possible to map AMR genes to higher level terms to test associations of higher level AMR classes with GIs.

4.2.2. Mobile genetic element prediction

To detect MGEs in the collection of genomes, GIs were predicted using IslandViewer, which includes predictions by three algorithms: IslandPath-DIMOB, SIGI-HMM, and IslandPick. GI predictions include such mobile elements as prophage, integrons, conjugative transposons, integrated plasmids, and ICEs. Plasmids not integrated into the chromosome were also considered as a separate group of MGEs. All genomes investigated had pre-computed GI predictions available. The GI dataset was sliced in various ways to test whether statistical significance held true under these different scenarios. For one, because GI boundary prediction is not perfect, a more inclusive GI dataset allowed AMR genes that are within 1000 base pairs (roughly the size of one gene) of a GI boundary to be considered part of the GI. A second dataset only focused on GI predictions by IslandPath-DIMOB, which has the highest precision and recall of sequence composition-based GI prediction methods, and a third set using IslandPath-DIMOB version 2 that greatly improves GI predictions from the original algorithm (unpublished). The dataset of AMR genes was combined with each of the different GI prediction datasets and plasmids to find AMR genes that overlapped with these MGEs to test statistical significance of any associations.

4.2.3. Statistical testing

Associations of AMR genes with various regions of the genomes were tested by first generating 2x2 contingency tables comparing counts of AMR genes inside and outside of GIs, plasmids, and the rest of the genome (for clarity, I will refer to the genomic regions of the chromosome outside of GIs as nonGIs). A Fisher's exact test was used to test significance. All statistical tests were performed using *R* and p-values smaller than 0.05 were considered to be significant. Secondly, for testing associations of higher level ARO terms with GIs, plasmids, and the rest of the genome, a similar approach was used, but the p-values were adjusted using the Benjamini and Hochberg False Discovery Rate correction for multiple testing.

4.3. Results and Discussion

4.3.1. Collectively, AMR is *not* disproportionately associated with mobile elements

Large-scale analysis of AMR genes and their association with GIs and/or plasmids in more than 2,700 bacterial and archaeal genomes reveals that AMR genes collectively are surprisingly not disproportionately associated with these mobile elements. Instead, AMR genes are found at higher, or similar (depending on the dataset of GI predictions used), levels in nonGI regions of the genome. A detailed list of raw counts of AMR genes overlapping with nonGIs, GIs, and plasmids for each tested GI dataset and p-values of significance when comparing AMR genes present inside versus outside of the specified regions is presented in Table 4.1. The first three rows of Table 4.1 show that when using the exact IslandViewer GI predictions for all MicrobeDB version 88 genomes, AMR genes are found at a significantly higher proportion in nonGIs than on GIs or plasmids (which don't have any significant difference). But when I use the more inclusive GI boundaries to include any AMR genes within 1000 base pairs of a GI boundary, then nonGIs and GIs have similar levels of AMR genes, that are significantly higher than the levels on plasmids. Both IslandPath-DIMOB GI datasets (which are more specific) display similar patterns; nonGIs have significantly higher proportions of AMR genes than GIs and plasmids. When analyzing the less-biased collection of genomes, which should reduce any sampling bias,

again there is a slightly higher proportion of AMR genes in nonGI regions, but the difference is not very significant due to the lower number of genes overall.

Table 4.1 AMR gene overlap with MGEs (GIs and plasmids) versus nonGI regions and significance for the various tested GI datasets

GI Dataset	Region	Total genes in region	Total AMR genes in region	% AMR genes in region	P-value of significance*
IslandViewer GIs with strict boundaries	NonGIs	7,587,023	25,770	0.340%	<2.20E-16
	GIs	562,708	1,514	0.269%	<2.20E-16
	Plasmids	256,989	717	0.279%	7.81E-07
IslandViewer GIs with 1000 bp inclusive boundaries†	NonGIs	7,587,023	25,391	0.335%	1.50E-02
	GIs	562,708	1,893	0.336%	6.49E-01
	Plasmids	256,989	717	0.279%	7.81E-07
IslandPath-DIMOB GIs with strict boundaries	NonGIs	7,737,939	26,042	0.337%	<2.20E-16
	GIs	411,792	1107	0.269%	1.51E-13
	Plasmids	256,989	717	0.279%	1.63E-06
IslandPath-DIMOB v2 GIs with strict boundaries	NonGIs	7,699,266	26,009	0.338%	<2.20E-16
	GIs	551,429	1,127	0.223%	<2.20E-16
	Plasmids	256,989	717	0.279%	5.31E-06
IslandViewer GIs with strict boundaries on less-biased genomes	NonGIs	418,186	1,431	0.342%	4.19E-02
	GIs	34,291	96	0.280%	6.53E-02
	Plasmids	11,835	35	0.296%	5.19E-01

* P-values of significance were calculated using a Fisher's Exact test comparing number of AMR genes inside versus outside of the specified regions

† Inclusive boundaries were applied to counting overlap of AMR genes with GIs and not raw counts of GI genes

So, when considering the bigger picture, what these values are describing is a situation that is very different than that of VFs. When analyzing VFs, they were extremely significantly associated with GIs, and this held true no matter which dataset of GIs or collection of genomes was used (Ho Sui et al., 2009). In the case of AMR genes we are able to see that, although many AMR genes are known to spread via HGT, collectively, microbial genomes tend to also intrinsically encode similar levels of AMR within the vertically inherited genome. This observation could be explained in part because we expect SNVs and other mutations that cause resistance to be located mainly on genes within nonGI regions. However, the levels of gene variants that cause resistance (as

measured by using the ARO) are not highly significantly disproportionately associated with nonGIs (see raw data presented in Table 4.5 of section 4.3.5) and on its own cannot explain these findings. Instead, the high levels of AMR genes on nonGIs suggests that the evolutionary pressures presented to AMR genes are different than those presented to VFs and may be contributing to the vertical inheritance and/or loss of mobility of mobile AMR genes. In Figure 4.1, panel (A) shows the acquisition of a toxin VF gene that, over time, can cause damage and even death of the host cell. Thus, such VFs do not confer any selective advantage and may actually be disadvantageous, so there is strong selective pressure for VFs like toxins to remain mobile and move in/out of genomes (Gill and Brinkman, 2011). On the other hand, panel (B) of Figure 4.1 highlights the case of acquiring an AMR gene that is required for survival when a particular antibiotic is pumped into the environment. Because of the strong selective pressure placed on microbes in the presence of antibiotics, it is essential for any progeny to gain such AMR genes for survival. What's more, progeny that do not retain these critical AMR genes will die, so there is strong selective pressure for any horizontally acquired AMR genes to also lose the ability to "pop out" of the genome. These differences in selective pressures placed on VFs and AMR genes could partially explain why AMR genes collectively, in contrast to VFs, are *not* disproportionately associated with GIs.

In addition to this, these findings also support the ancient origin theory of AMR (D'Costa et al., 2006; Finley et al., 2013). Anciently acquired AMR genes would have benefitted microbes in the environment that were exposed to antibiotics regularly and instead of moving in and out of genomes, they would have strong selective pressure to be retained and shared vertically with new generations, as mentioned previously. If these ancient AMR genes do indeed have foreign origins, it is plausible that over time these genes have ameliorated to match sequence composition of the host genome and could have lost associations with mobile elements, as they became stable components of the genome of these specific organisms. Overall, high levels of AMR genes within nonGIs is a result of the different selective pressures that are presented on AMR genes in contrast to VFs and could also provide evidence to support the ancient origin of AMR.

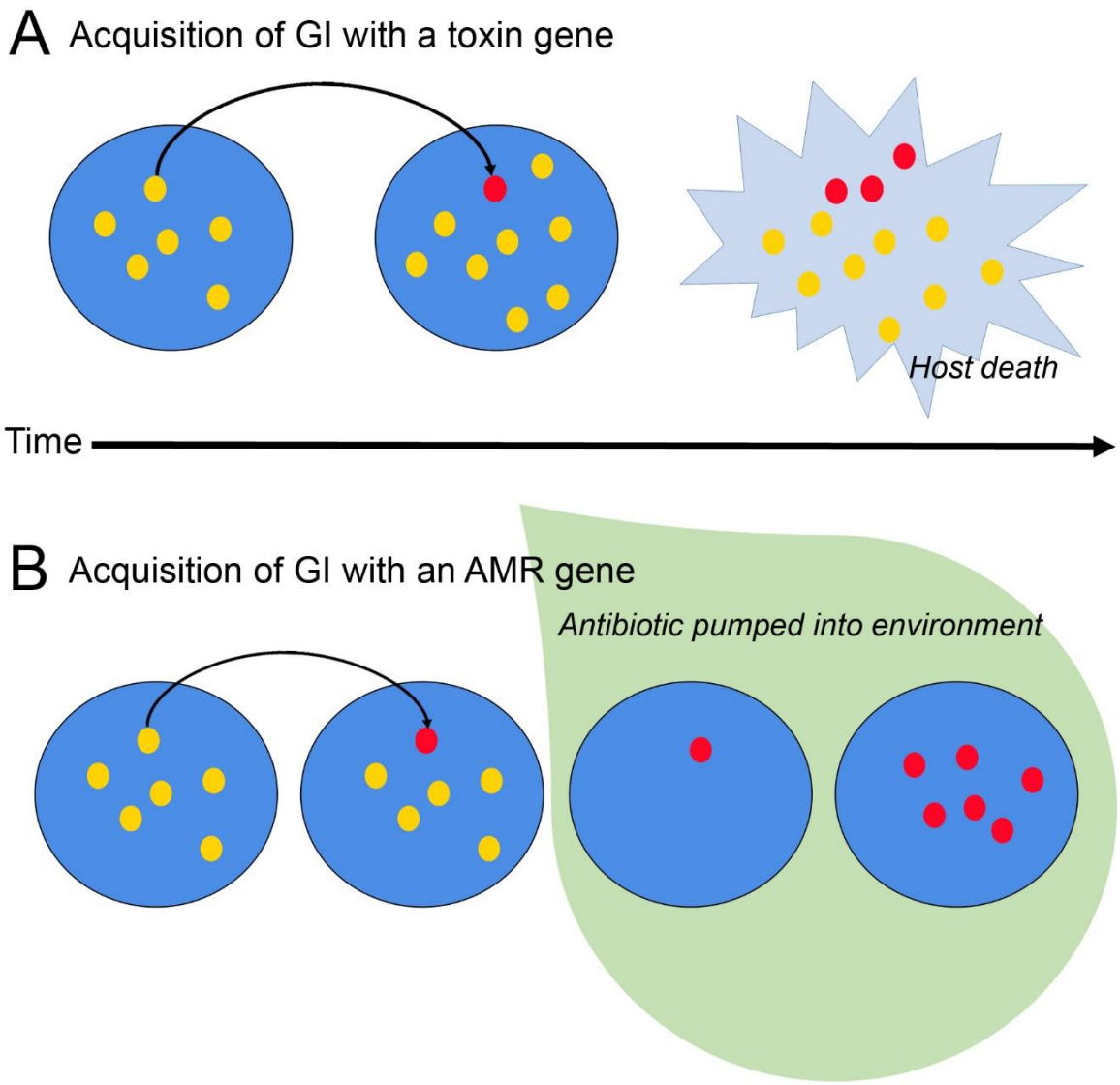

Figure 4.1 Difference in selective pressures presented to VFs and AMR genes acquired on GIs

In this example, the blue circles represent a host cell, the yellow circles represent microbes living within the host, and the red circles depict microbes that acquired a GI encoding either (A) a toxin VF gene which over time destroys the host cell, or (B) an AMR gene which is shown to be necessary for survival in the presence of an antibiotic being pumped into the environment.

The analysis presented in this section is important in gaining perspective on global trends of mobility of AMR genes across diverse genomes, however, there are some caveats to such an approach. One important factor to bear in mind is that our dataset is biased towards known AMR genes and does not represent any novel undiscovered AMR genes. Thus, there may be under-represented classes of AMR genes in our dataset.

Furthermore, this analysis is also biased by the collection of genomes that are available. Although I made attempts to reduce the sampling bias of genomes, it is not known whether this is a true representation of the natural microbial population. Nonetheless, this analysis provides insight into trends of AMR genes and the evolution of GIs over this collection of genomes that can be tested further as more genomes become available. Also by grouping all AMR genes together in one category, it does not capture the distinctive trends that may be present for the numerous resistance classes represented in this dataset. So I further investigated the trends of high-level AMR classes by using the ARO terms associated with each AMR gene to map to higher levels in the ontological hierarchy. I performed these tests using the GI datasets from IslandViewer with strict boundaries and all MicrobeDB version 88 genomes.

4.3.2. Certain AMR classes are disproportionately associated with mobile elements

By using the ARO to classify AMR genes into higher-level resistance classes, I uncovered bias in classes that have disproportionately more AMR genes on mobile elements than the rest of the genome, and vice versa. Figure 4.2 summarizes the overall trends in proportions of AMR genes on nonGIs, GIs, and plasmids from the various high-level classes. This figure clearly shows that the various classes have different trends in association with mobile elements. There are clusters of classes that are found at higher proportions on GIs, plasmids, or both and will be discussed below, while others are found at higher levels on nonGIs and will be discussed in more detail in 4.3.3. There are also classes that have no significant association with any region and will be discussed in more detail in section 4.3.4.

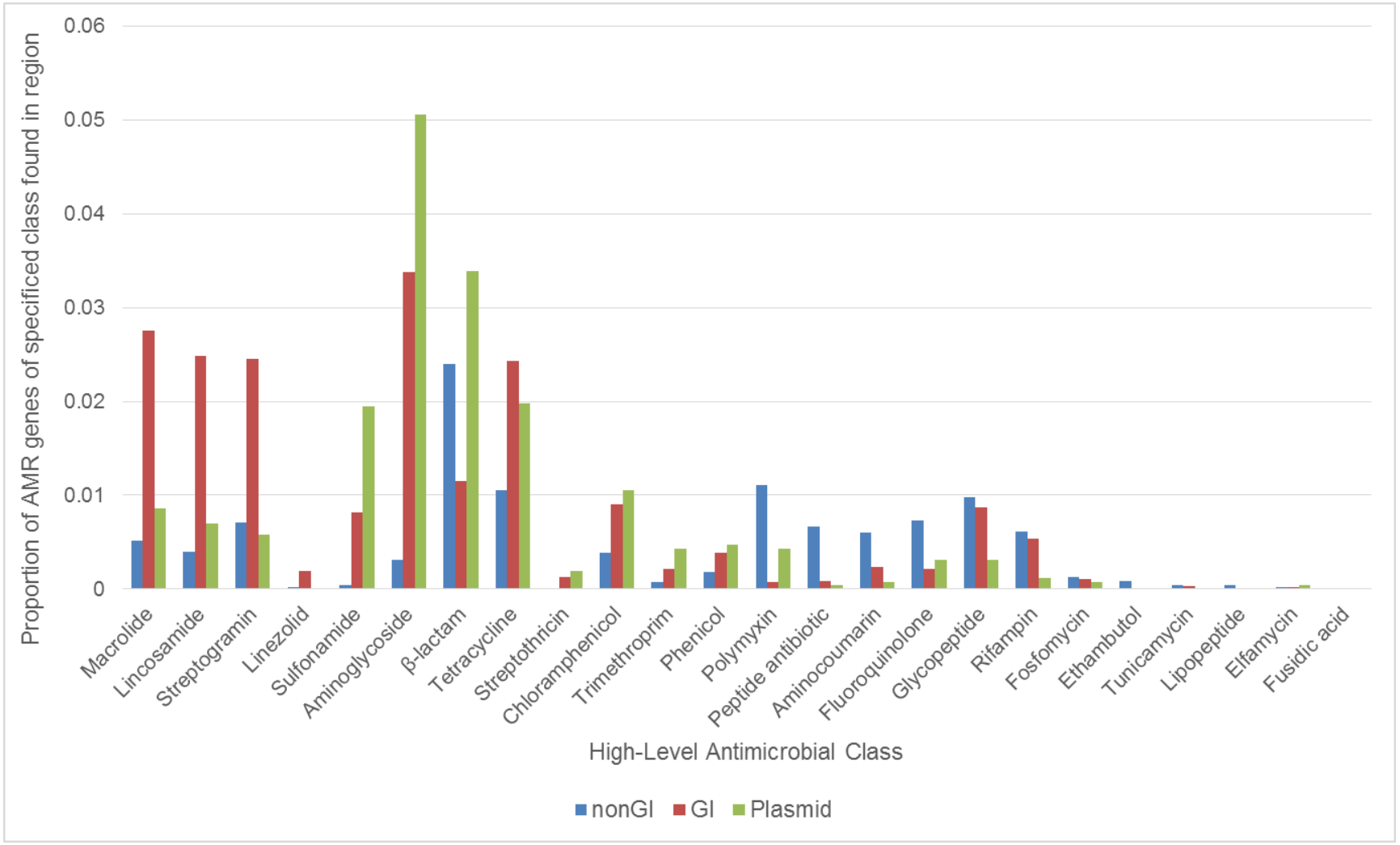

Figure 4.2 Proportions of AMR genes broken down by high-level classes found on nonGIs, GIs, and plasmids.

Macrolide, lincosamide, streptogramin, and linezolid resistance genes are found at at significantly higher levels on GIs, while sulfonamides, aminoglycosides, and β-lactams AMR genes are higher on plasmids, and tetracyclines, streptothricin, phenicol, chloramphenicol, and trimethoprim are found at similar levels on GIs and plasmids that are significantly higher than levels on nonGIs. The raw counts and proportions of AMR genes from each of these classes that are significantly associated with mobile elements are presented in

Table 4.2. Although previous studies have shown mobile elements play a role in the spread of these resistance classes, this study reveals that GIs and/or plasmids are the primary method of spread of resistance for these classes and they have not integrated into nonGI regions at comparable levels.

Table 4.2 Summary of high-level AMR classes with significant associations to MGEs (i.e. GIs and/or plasmids)

ARO term	AMR genes on nonGIs (%)	AMR genes on GIs (%)	AMR genes on plasmids (%)	Associated with (p-value*)
Macrolide ARO:3000315	389 (0.005%)	155 (0.028%)	22 (0.009%)	GIs (<2.2e-16)
Lincosamide ARO:3000240	302 (0.004%)	140 (0.025%)	18 (0.007%)	GIs (<2.2e-16)
Streptogramin ARO:3000241	541 (0.007%)	138 (0.025%)	15 (0.006%)	GIs (<2.2e-16)
Linezolid ARO:3000267	19 (0.000%)	11 (0.002%)	0 (0.000%)	GIs (2.0e-06)
Sulfonamide ARO:3000408	32 (0.000%)	46 (0.008%)	50 (0.019%)	Plasmids (<2.2e-16)
Aminoglycoside ARO:3000104	237 (0.003%)	190 (0.034%)	130 (0.051%)	Plasmids (<2.2e-16)
β-lactam ARO:3000129	1,820 (0.024%)	65 (0.012%)	87 (0.034%)	Plasmids (1.0e-03)
Tetracycline ARO:3000472	797 (0.011%)	137 (0.024%)	51 (0.020%)	GIs & plasmids (<2.2e-16)
Streptothricin ARO:3000868	0 (0.000%)	7 (0.001%)	5 (0.002%)	GIs & plasmids (7.4e-13)
Chloramphenicol ARO:3000398	289 (0.004%)	51 (0.009%)	27 (0.011%)	GIs & plasmids (4.9e-11)
Trimethroprim ARO:3001217	59 (0.001%)	12 (0.002%)	11 (0.004%)	GIs & plasmids (2.4e-06)
Phenicol ARO:3000052	136 (0.002%)	22 (0.004%)	12 (0.005%)	GIs & plasmids (4.7e-05)

* P-value calculated using Fisher's Exact test and adjusted for multiple testing using the Benjamini and Hochberg False Discovery Rate

GI-associated AMR classes

The *erm* (erythromycin resistance methylase) genes are broad acting AMR genes capable of conferring resistance against three classes of antibiotics: macrolide-lincosamide-streptogramin (MLS) (Lina et al., 1999). The MLS antibiotics are structurally dissimilar but all target protein synthesis by binding the 23s rRNA subunit of the 50S ribosome. *Erm* genes can methylate the 23s subunit to prevent binding of the drug to its target. These genes are known to spread via transposons and sometimes plasmids

(Leclercq and Courvalin, 1991; Leclercq, 2002; Lina et al., 1999) and account for the high levels of MLS resistance genes found on predicted GIs versus the rest of the chromosome and even plasmids.

Linezolid is a synthetic antibiotic and thus had low chances of any naturally occurring resistance genes (Meka and Gold, 2004).. Studies on existing mechanisms of resistance against drugs similar to linezolid that inhibit ribosomal protein synthesis did not confer resistance against linezolid (Fines and Leclercq, 2000). However, *in vitro* and clinical cases of disease with resistance against linezolid have been discovered and involve mutations in the target 23S rRNA as the mechanism of resistance (Meka et al., 2004; Prystowsky et al., 2001; Tsiodras et al., 2001; Zurenko et al., 1996). More recent reports have demonstrated that linezolid resistance can also be mediated by the naturally-occurring *cfr* (chloramphenicol-florfenicol resistance) gene (Mendes et al., 2008; Morales et al., 2010; Toh et al., 2007), which is known to be horizontally transferred via IS elements. This AMR gene was detected at higher levels within GIs than plasmids or nonGIs in our dataset and explains why linezolid resistance is detected as being significantly associated with GIs.

Plasmid-associated AMR classes

Sulfonamides act by competitively inhibiting the enzyme dihydropteroate synthase (*dhps* or *folP*) that plays a role in folate synthesis and results in growth inhibition of bacteria. Mutations in the target gene have been shown to cause resistance in many different species (Gibreel and Sköld, 1999; Helweg-Larsen et al., 1999; Kai et al., 1999; Kristiansen et al., 1990; Lopez et al., 1987; Mei et al., 1998; Radstrom et al., 1992; Swedberg et al., 1998). Sulfonamide resistance genes *sul1* and *sul2* both encode a resistant *dhps* and have long been known to be mobile and transmitted via plasmids (Sköld, 1976; Wise and Abou-Donia, 1975), but they have also been linked to other MGEs such as integrons and non-conjugative plasmids (Rådström et al., 1991; Sköld, 2000; van Treeck et al., 1981). Foreign copies of resistant *folP* can also be transmitted horizontally (Fermer et al., 1995; Sköld, 2000; Sköld, 1976; Wise and Abou-Donia, 1975). Altogether, the mobile *sul1*, *sul2*, and *folP* AMR genes account for the disproportionately higher levels of sulfomanide resistance genes on plasmids.

Aminoglycosides target ribosomal proteins to inhibit protein synthesis. Resistance can result from multiple mechanisms including enzymatic modification of the drug for inactivation, blocking entry or removing drug, methylation of aminoglycoside binding sites, or alteration of ribosomal target proteins through mutation (Shakil et al., 2008). AMR genes against aminoglycosides have been previously found to be associated with mobile elements, mostly plasmids, but also integrons and transposons, to be commonly transferred via HGT (Courvalin and Calier, 1981; Doi and Arakawa, 2007; Hall and Collis, 1998; Mingeot-Leclercq et al., 1999; Nemec et al., 2004; Shakil et al., 2008; Vakulenko et al., 2003; Zarrilli et al., 2005). Enzymes from all major classes of aminoglycoside-modifying enzymes (acetylation, adenylylation, and phosphorylation) were well represented in this dataset and contributed to the significant association of aminoglycoside resistance with plasmids.

Lastly, resistance genes against β-lactams are found across all regions of the genome, but at significantly higher proportions on plasmids than GIs or nonGIs. β-lactams inhibit the synthesis of peptidoglycan by binding proteins called penicillin-binding proteins (PBPs) and preventing them from forming cross-links between cell wall components. β-lactamases can inhibit the activity of β-lactams by altering its structure to prevent binding to PBPs. There are a variety of different β-lactamase enzymes including the OXA-type, TEM-type, *ampC*-type, CTX-M-type, *bla*-type among others that have been identified on plasmids (Bonnet, 2004; Knox, 1995; Murray and Mederski-Samaroj, 1983; Papanicolaou et al., 1990; Philippon et al., 2002) and some of which are known to be associated with integrons and transposable elements (Hall and Collis, 1998; Naas et al., 1998; Ouellette et al., 1987; Sidhu et al., 2002). Another mechanism of resistance to β-lactams is through mutation of PBPs that prevents or lowers affinity of binding of the β-lactams (Hartman and Tomasz, 1984; Malouin and Bryan, 1986; Zapun et al., 2008). These AMR genes were also observed in our dataset and a large number of these mutated PBPs were found on nonGIs. “Intrinsic resistance” against β-lactams has also been observed, but is likely due to broader resistance mechanisms such as efflux pumps (Sanders and Sanders, 1983) that can affect a large range of different antibiotics. In all, resistance against β-lactam drugs is known to be largely spread via plasmids, but can also be chromosomally encoded and this analysis supports these trend in that AMR genes against β-lactams are largely associated with plasmids. However, the significance is not as much as other mobility-

associated classes since there are also a large number of AMR genes found on nonGIs (at levels higher than GIs).

GI and plasmid-associated AMR classes

Tetracyclines can actively inhibit protein synthesis by binding to ribosomes or destroy the cell membrane by binding to phospholipid or protein components of the membrane (Schnappinger and Hillen, 1996). Resistance is primarily mediated by efflux pumps, ribosomal protection proteins, or modification of the drug and a large majority of these genes are known to be associated with mobile elements (Roberts, 1996). Gram-negative efflux genes are typically found on transposons inserted on plasmids (Jones et al., 1992; Mendez et al., 1980; Roberts, 1996), while Gram-positive efflux genes are found on small mobilizable plasmids (Khan and Novick, 1983; Schwarz et al., 1992). Similarly, ribosomal protection proteins have been found on plasmids and conjugative transposons (Charpentier et al., 1993; Li et al., 1995; Naglich and Andrews, 1988; Salyers et al., 1995a; Taylor and Courvalin, 1988; Torres et al., 1991). This supports the finding that tetracycline resistance genes are found on plasmids and GIs at significantly higher proportions than nonGIs.

Streptothricin is another protein synthesis inhibitor that functions by binding to ribosomes. Known resistance mechanisms, however, are not as common and only include drug modification enzymes (acetyltransferases and adenyltransferases) that have been isolated from streptothricin-producing organisms and some clinical pathogens (Hamano et al., 2006; Kobayashi et al., 1986). These resistance genes are also found associated with transposons and integrons (Partridge and Hall, 2005; Peirano et al., 2005; Singh et al., 2005). The low abundance of streptothricin resistance genes overall in this dataset could be attributed to the fact that it has not been approved for clinical use in many countries due to its high toxicity and resistance has not spread as much as other AMR classes. Nonetheless, this drug has resistance genes that are associated with mobile elements that could allow for widespread distribution.

Chloramphenicol is a part of the phenicol family of antibiotics and should be classified under the phenicol ARO term. This is an error in the structure of the ontology. The phenicols are yet another class of protein synthesis inhibitors functioning through

binding of ribosomes. Modification of phenicol drugs is a common mechanism of resistance encoded by acetyltransferases (e.g. *cat* chloramphenicol acetyltransferase) found on various plasmids and transposons (Alton and Vapnek, 1979; Horinouchi and Weisblum, 1982; Shaw and Brodsky, 1968). Another gene, *cfr* (chloramphenicol-florfenicol resistance), is known to be transmitted via integrons on plasmids (Kehrenberg et al., 2007; Long et al., 2006). Efflux pumps that act against phenicol drugs, including *floR*, are commonly seen on transposons and integrons (Hall and Collis, 1998). As such, it is not surprising that this analysis supports a significant association of phenicol resistance genes with GIs and plasmids.

Trimethoprim is a drug that binds dihydrofolate reductase (*dfr*) to ultimately inhibit DNA synthesis. The most commonly observed resistance mechanism involves the use of an alternative *dfr* gene that is resistant to binding by trimethoprim and is found on transposons, integrons and plasmids (Blahna et al., 2006; Charpentier and Courvalin, 1997; Datta and Hedges, 1972; Hall and Collis, 1998; Rouch et al., 1989; Shapiro and Sporn, 1977; Skold and Widh, 1974; Threlfall et al., 1980; Towner et al., 1980). This gene accounts for the disproportionate association of trimethoprim resistance genes with GIs and plasmids.

Overall trends of mobility-associated AMR classes

As described in this section, this analysis revealed a collection of AMR classes that are disproportionately associated with mobile elements, specifically GIs and plasmids. For one thing, many of the AMR classes associated with GIs and/or plasmids target ribosomal proteins to inhibit protein synthesis (MLS, linezolid, aminoglycosides, tetracyclines, streptothricin, phenicols). These types of antimicrobial agents may be particularly vulnerable to the acquisition of mobile AMR genes since chromosomal mutations in binding pockets of ribosomal proteins may be detrimental to the organism in other ways. As such, development of resistance mutations is a delicate balance between reducing or removing interaction with antibiotics and preserving the function of the protein, and for antibiotics targeting very critical regions of proteins, the use of novel genes gained from foreign sources is likely an important source of resistance, as is observed in the case of AMR against antimicrobials that target ribosomes. In addition, it is evident from this study that the horizontal spread of AMR genes is not limited to naturally-occurring

antibiotics based on the observation that AMR genes against synthetically-derived antibiotics such as linezolid, sulfonamide, and trimethoprim also have higher proportions on GIs and/or plasmids. This is an important consideration for the future development of antimicrobial agents because this shows that if there is strong enough selective pressure present, resistance can emerge even against synthetically developed drugs.

4.3.3. Certain AMR classes are *not* disproportionately associated with mobile elements

On the contrary, I also reveal classes of resistance that do not have any association with mobile elements and are mediated mainly by chromosomal mutations. These include polymyxins, peptide antibiotics, aminocoumarins, and fluoroquinolones (see Table 4.3). Other lipopeptide antibiotics and ethambutol resistance classes did not ever have AMR genes seen on GIs or plasmids, but the number of AMR genes for members of these classes was simply too low for a statistically significant result.

Table 4.3 Summary of higher-level AMR classes not associated with MGEs

ARO term	AMR genes on nonGIs (%)	AMR genes on GIs and plasmids (%)	P-value of significance*
Polymyxin ARO:3002984	842 (0.011%)	15 (0.002%)	<2.2e-16
Peptide antibiotic ARO:3000751	503 (0.007%)	6 (0.001%)	1.5e-15
Aminocoumarin ARO:3000477	460 (0.006%)	15 (0.002%)	4.5e-08
Fluoroquinolone ARO:3000102	554 (0.007%)	20 (0.002%)	1.9e-08

* P-value calculated using Fisher's Exact test and adjusted for multiple testing using the Benjamini and Hochberg False Discovery Rate

Notably, the peptide and lipopeptide antibiotics (including polymyxin) and ethambutol all target and disrupt bacterial cell membranes or cell walls. Polymyxins specifically target lipopolysaccharide (LPS) molecules to disrupt the inner and outer cell membranes of Gram-negative bacteria (Landman et al., 2008; Newton, 1956). Resistance against polymyxins is acquired mainly through chromosomal mutations, including LPS modifications, formation of capsules, over-expression of outer membrane protein OprH,

and efflux pumps (Gunn et al., 1998; Lee et al., 2004; Moore and Hancock, 1986; Olaitan et al., 2014). Recently, a plasmid-mediated resistance gene, *mcr-1*, active against colisitin (a type of polymyxin) was discovered in China (Liu et al., 2015). Follow up studies revealed this plasmid is actually present across the globe in many different species, mostly *E. coli*, and is capable of HGT (Skov and Monnet, 2016). This newly discovered AMR gene was not represented in the CARD database when our analysis was conducted and thus, our results are biased towards the previously discovered AMR mechanisms. The recent discovery of *mcr-1* could skew polymyxins to have less of an association with nonGIs if this analysis were performed again with genomes harbouring *mcr-1*. Nonetheless, resistance against other peptide antibiotics is mainly seen as chromosomal mutations. For example, resistance against viomycin results from mutation of the ribosomal genes (Taniguchi et al., 1997; Yamada et al., 1972), and resistance against daptomycin is caused by mutations in *liaF* and *gdpD*-family proteins (Arias et al., 2011) that are targeted by these drugs. In all, resistance against peptide and lipopeptide antibiotics is mainly associated with chromosomal mutations within nonGI regions, thus resistance to these types of drugs could be considered low risk for spreading globally through HGT. However, as shown for colistin, it is possible that mobile AMR genes against these groups of drugs have just not been discovered yet and are under-represented in our dataset.

On the other hand, aminocoumarins and fluoroquinolones are interesting because they both bind DNA gyrase/topoisomerase (*gyr* or *par* family of genes) to inhibit proper unwinding and replication of DNA during cell division. The only known resistance mechanisms against these drug classes are mutations specifically in the gyrase/topoisomerase gene (Hooper, 2001; Schmutz et al., 2003). Some multidrug efflux pumps are also capable of removing fluoroquinolones from the cell (Hooper, 2001; Kaatz et al., 1993). Cases of plasmid-mediated resistance against fluoroquinolones and other quinolones have been discovered but the prevalence is not known (Martínez-Martínez et al., 1998; Martínez-Martínez et al., 2014). In this dataset, plasmid-mediated resistance to fluoroquinolones is found at lower levels than chromosomally-encoded resistance. Mainly, resistance against aminocoumarins and fluoroquinolones has been the result of chromosomal mutation of their targets and suggests that the risk of spread of resistance against these classes to other bacteria could be relatively low based on this analysis.

4.3.4. Other AMR classes have no associations with a particular region

Some AMR classes showed no specific associations with GIs, plasmids or non-GIs because either they are found more ubiquitously around these various regions of the genome, or there were too few AMR genes of the class in our dataset to find any statistically significant association. For those AMR genes that were found more ubiquitously, this may suggest there is no clear selection for such genes to be vertically or horizontally acquired and both mechanisms play significant roles in the spread of these resistance classes. The numbers and proportions of AMR genes from each of these classes in nonGIs, GIs and plasmids are summarized in Table 4.4. Also, a set of AMR classes present in the ARO were not detected in any of the analyzed genomes and warrant further investigation for their association with mobile elements in microbial genomes; this includes mupirocin, isoniazid, pyrazinamide, and polyamine resistances.

Table 4.4 AMR classes with no significant associations with GI, plasmids, or nonGIs

ARO term	AMR genes on nonGIs (%)	AMR genes on GIs (%)	AMR genes on plasmids (%)
Glycopeptide ARO:3000494	745 (0.010%)	49 (0.009%)	8 (0.003%)
Rifampin ARO:3000383	463 (0.006%)	30 (0.005%)	3 (0.001%)
Fosfomycin ARO:3000271	97 (0.001%)	6 (0.001%)	2 (0.001%)
Ethambutol ARO:3000468	63 (0.001%)	0 (0.000%)	0 (0.000%)
Tunicamycin ARO:3003058	35 (0.000%)	2 (0.000%)	0 (0.000%)
Lipopeptide ARO:3003073	32 (0.000%)	0 (0.000%)	0 (0.000%)
Elfamycin ARO:3001311	15 (0.000%)	1 (0.000%)	1 (0.000%)
Fusidic acid ARO:3003024	1 (0.000%)	0 (0.000%)	0 (0.000%)

Glycopeptides, such as vancomycin, inhibit peptidoglycan synthesis by binding to amino acids in the cell wall. Resistance genes (specifically from the *van* family) can

synthesize an alternative precursor for peptidoglycan polymerization that glycopeptides have much lower affinity for binding (Arthur et al., 1996; Arthur and Courvalin, 1993). Some of these resistance genes are found encoded intrinsically on the chromosome, such as *vanC*, while others are found associated with transposons, such as *vanA*, or plasmids, such as *vanB* (Courvalin, 2006). Because of this, glycopeptide resistance genes are identified at similar proportions on GIs and nonGIs and are not disproportionately associated with mobile elements.

Rifampin is a drug that that binds RNA polymerase (specifically the β subunit encoded by *rpoB*) to effectively inhibit DNA transcription. Resistance against rifampin can be caused by chromosomal mutations in *rpoB* (Aubry-Damon et al., 1998; Heep et al., 1999; Wehrli, 1983). Rifampin can also be inactivated by ribosylation (encoded by *arr*), glycosylation (encoded by *rgt*), phosphorylation, and decomposition, some of which are encoded by AMR genes found on integrons (Dabbs et al., 1995a; Dabbs et al., 1995b; Quan et al., 1997; Spanogiannopoulos et al., 2012; Tanaka et al., 1996; Yazawa et al., 1993; Yazawa et al., 1994). Efflux pumps encoded on plasmids have also been implicated in rifampin resistance (Chandrasekaran and Lalithakumari, 1998) and are seen at low levels in this dataset. Thus, AMR genes against rifampin are detected at similar levels across the chromosome, inside or outside of GIs in this analysis.

Fosfomycin is capable of blocking peptidoglycan synthesis by modifying *murA* that is a critical protein in this process. Resistance can be mediated via chromosomal mutation of *murA* to prevent inactivation (Kim et al., 1996), or modification of fosfomycin by the *fosA* gene that has been seen on conjugative transposons (Arca et al., 1988; Garcia-Lobo and Ortiz, 1982) and the proportions of these resistance genes are similar across GIs, plasmids and nonGIs.

4.3.5. Particular mechanisms of action of AMR genes are associated with mobile elements or nonGIs

I also investigated associations between mechanisms of action of AMR genes and mobile regions of the genome (see results summarized in Figure 4.3 and raw counts and p-values of significance presented in Table 4.5). This type of analysis is important in

gaining a better understanding of trends in the types of over-arching mechanisms of resistance that are more likely to result in the development of mobile AMR genes.

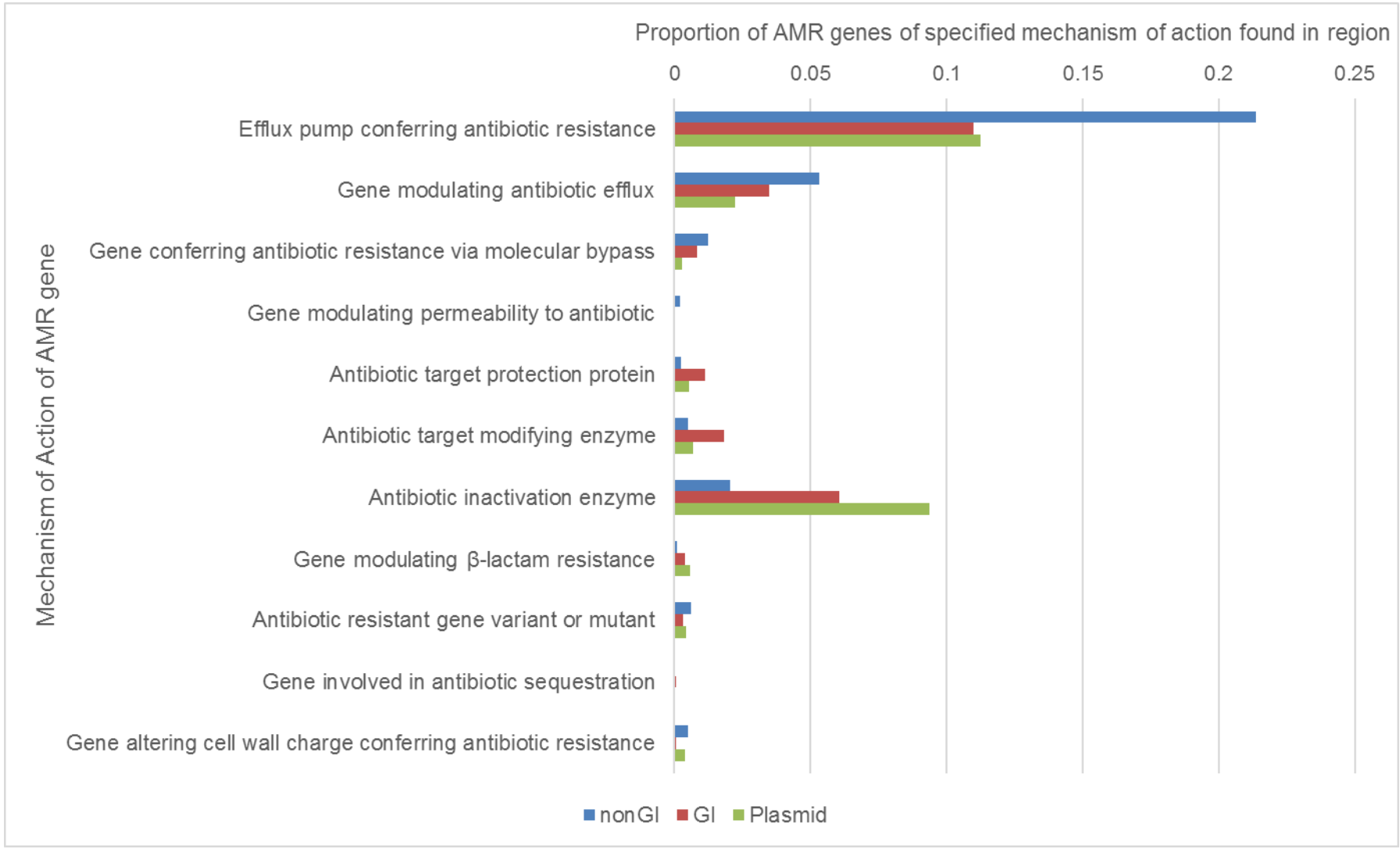

Figure 4.3 Proportions of AMR genes broken down by mechanism of action found on nonGIs, GIs, and plasmids

First, efflux pumps and genes modulating efflux to actively remove multiple different types of antibiotic molecules are found at disproportionately higher levels in nonGI regions. These have previously been known to be chromosomally-encoded rather than spread via mobile elements, are even found in some eukaryotes and archaea (Martinez et al., 2009; Poole, 2005), and may be considered more ancient mechanisms of resistance that act broadly against a range of toxic agents. Genes conferring antibiotic resistance via molecular bypass are those involved in restructuring of the cell wall and include the *van* proteins involved in glycopeptide resistance (as discussed previously in section 4.3.4) and are also disproportionately associated with nonGIs. In addition, genes modulating permeability to antibiotics are found at significantly higher levels on nonGIs. In this dataset, the *mar* (multiple antibiotic resistance) genes were the main contributors to this mechanism and were only seen on nonGIs. *mar* genes participate in changing membrane permeability in order to prevent the uptake of a broad range of antibiotics (Cohen et al., 1989; Cohen et al., 1993; Moken et al., 1997). Taken together, the mechanisms of action that are disproportionately associated with nonGIs tend to be broad-acting and provide resistance against multiple classes of antimicrobials. These mechanisms could be more anciently developed before specialized systems targeting specific antibiotics became widespread and have become integral components of the heritable genome for certain organisms.

On the other hand, other mechanisms are seen at higher proportions on GIs and/or plasmids. This includes antibiotic target protection proteins (e.g. the *tetM* ribosome protection protein that prevents tetracycline from binding), target modifying enzymes (e.g. *ermA* that methylates 23S rRNA for resistance against erythromycin), antibiotic inactivation enzymes (e.g. β-lactamases that break down β-lactams), and genes modulating β-lactam resistance (e.g. *blaR*). This set of mechanisms overall represent more specialized techniques that target distinctive molecules and may have evolved later than more ancient AMR genes that act more broadly. Such AMR mechanisms may also confer some advantage of being horizontally transmitted, for example there may be a high cost to maintain such AMR mechanisms in the chromosome, but this issue warrants further study. Overall, the trends observed in this part of the analysis indicate there are differences in the mobility of certain resistance mechanisms, and this information could be useful for future development of antimicrobials.

Table 4.5 Associations of mechanisms of action of AMR genes

ARO term	AMR genes on nonGIs (%)	AMR genes on GIs (%)	AMR genes on plasmids (%)	Associated with (p-value*)
Efflux pump conferring antibiotic resistance ARO:3000159	16,205 (0.213%)	618 (0.110%)	289 (0.112%)	NonGIs (<2.2e-16)
Gene modulating antibiotic efflux ARO:3000451	4,030 (0.053%)	196 (0.035%)	57 (0.022%)	NonGIs (<2.2e-16)
Gene conferring antibiotic resistance via molecular bypass ARO:3000012	947 (0.012%)	46 (0.008%)	7 (0.003%)	NonGIs (3.6e-07)
Gene modulating permeability to antibiotic ARO:3000270	158 (0.002%)	1 (0.000%)	0 (0.000%)	NonGIs (2.8e-06)
Antibiotic target protection protein ARO:3000185	171 (0.002%)	64 (0.011%)	14 (0.005%)	GIs (<2.2e-16)
Antibiotic target modifying enzyme ARO:3000519	374 (0.005%)	103 (0.018%)	18 (0.007%)	GIs (<2.2e-16)
Antibiotic inactivation enzyme ARO:3000557	1,550 (0.020%)	340 (0.060%)	241 (0.094%)	Plasmids (<2.2e-16)
Gene modulating β-lactam resistance ARO:3000100	61 (0.001%)	21 (0.004%)	15 (0.006%)	GIs & Plasmids (4.8e-13)
Antibiotic resistant gene variant or mutant ARO:0000031	452 (0.006%)	17 (0.003%)	11 (0.004%)	None
Gene involved in antibiotic sequestration ARO:3001207	6 (0.000%)	3 (0.001%)	1 (0.000%)	None
Gene altering cell wall charge conferring antibiotic resistance ARO:3003580	384 (0.005%)	3 (0.001%)	10 (0.004%)	None

* P-value calculated using Fisher's Exact test and adjusted for multiple testing using the Benjamini and Hochberg False Discovery Rate

4.4. Conclusions

In summary, this is the first large-scale study across all bacterial genomes sequenced to date, revealing that AMR genes collectively, unlike VFs, are not disproportionately associated with mobile elements like GIs and plasmids, and rather are found at higher (or similar) levels in nonGI regions. This is due in part because of chromosomal variants linked to resistance, but may also be a result of vertical inheritance of AMR genes that are under high selective pressure to remain stable in the genome and lose mobility components. This is unlike VFs that are under considerable selective pressure to remain mobile and move in/out of genomes. This analysis could also be evidence supporting the ancient origin of AMR in that older AMR genes have incorporated and become integral components of certain genomes to be vertically passed on to progeny. These ancient AMR genes could also have ameliorated to match the host genome and lost the association with mobile elements and can no longer be detected as genes of foreign origin. By breaking down the dataset further into mechanisms of action of the AMR genes using the ARO, I was able to show that there are mechanisms of action that are disproportionately associated with nonGIs that also support ancient origin because they are broad and wide-ranging in the classes of antimicrobial agents they provide protection against. These include multidrug efflux pumps, genes modulating cell permeability, and genes involved in restructuring of the cell wall. These types of AMR genes could have appeared in microbial life before specialized AMR genes that target very distinctive molecules and therefore support origins that are more ancient.

I was also able to break down the AMR genes dataset by high-level resistance classes using the ARO. Notably, there are high-level classes of AMR genes that are disproportionately associated with nonGIs, such as polymyxins, peptide antibiotics, aminocoumarins and fluoroquinolones. However, these associations do not suggest ancient origins of these AMR genes. Instead, these are significantly associated with nonGIs because known resistance mechanisms against these classes involve mutation of chromosomal target genes. There may be undiscovered or recently emerging mechanisms that are spread via mobile elements that are not represented in our dataset, such as the newly discovered plasmid-borne *mcr-1* gene for colistin resistance. Nonetheless, these nonGI-associated AMR classes represent resistances that may be

considered lower risk for rapid spread globally because they are not associated with mobile elements at significant levels.

On the contrary, this study has also revealed classes of resistance that are associated with mobile elements, specifically GIs and/or plasmids, that may be considered higher risk of spreading globally between diverse genera. Many of these AMR genes have already spread to unprecedented levels and threaten the ability to use certain front-line drugs against infections. This includes macrolides-lincosamides-streptogramins (MLS), linezolid, sulfonamides, aminoglycosides, β-lactams, tetracyclines, streptothricin, phenicols, and trimethoprim. This represents a large majority of AMR classes and reinforces that GIs and plasmids play an important role in the spread of AMR genes. These data also demonstrate that very few AMR classes are not affected by the development of mobile AMR genes. Even the few AMR classes with AMR genes that were detected at higher levels on nonGIs may also become more associated with mobile elements in the near future upon discovery of new resistance mechanisms that are able to spread via HGT.

In all, this chapter presents the first large-scale analysis of the mobility of AMR genes by combining datasets of AMR genes and GIs from diverse microbial organisms to reveal important trends that are worth further study. This study provides evidence in support of the ancient origin of resistance, can help inform risk assessment of classes of AMR capable of spreading via mobile elements, and also provides important insight into the evolution of GIs in the context of AMR.

Chapter 5.

Application of IslandViewer to *Listeria monocytogenes* food-borne outbreaks

5.1. Introduction

As described in the introductory section 1.5, *L. monocytogenes* is an important food-borne pathogen that causes listeriosis with high mortality rates. It was first recognized as a food-borne pathogen linked to contaminated coleslaw in 1981 in Canada (Schlech III et al., 1983) and then pasteurized milk in 1983 in the United States (Fleming et al., 1985). Outbreaks of *L. monocytogenes* have since been linked to a variety of different food vectors including cantaloupe, celery, and especially cold ready-to-eat foods like soft cheeses and deli meats (Cartwright et al., 2013). Recent studies have begun to use WGS to better understand this important pathogen's genomic structure to reveal that *L. monocytogenes* genomes are essentially syntenic and the major differences lie in SNVs and GIs (especially phage) (Gilmour et al., 2010; Nelson et al., 2004). It is becoming increasingly important to not only study SNVs, but also mobile genetic elements such as GIs in genomic investigations of outbreaks. A recent study on *L. monocytogenes* outbreaks in Australia highlighted the importance of using GIs to distinguish between outbreak-related cases and unrelated sporadic cases (Wang et al., 2015). Previous studies have also shown that strains of *L. monocytogenes* collected from the same location 12 years apart have very little variation in backbone genome, but significant differences in phage evolution (Orsi et al., 2008). More recent studies have shown clones from within a single *L. monocytogenes* outbreak can also have significant variation within phage regions (Bergholz et al., 2015) and other mobile elements that could help distinguish outbreak-related cases from unrelated sporadic cases (Wang et al., 2015). Overall, the *L. monocytogenes* pan-genome is not closed as there is considerable variation in the accessory genome including a wide array of horizontally acquired DNA, especially phage (Klumpp and Loessner, 2013).

Because of these characteristics, *L. monocytogenes* is a reasonable model organism to use for studying the dynamics of GIs between and during outbreaks. I had access to a relatively large dataset of *L. monocytogenes* genomes from the Public Health Agency of Canada that represent diverse outbreaks from across Canada to perform such an analysis. These genomes were sequenced by the Public Health Agency as an initial evalution of using WGS to determine genetic variability of epidemiologically related and sporadic isolates and focused mainly on SNV differences (Aleisha Reimer *et al.*, unpublished). This chapter will focus on applying IslandViewer's GI prediction pipeline on *L. monocytogenes* genomes collected from various food-borne outbreaks from across Canada over a period of 30 years to show that GI content differs between each distinct outbreak. I have also used our dataset to investigate the dynamics and stability of GIs within a single large outbreak to further examine whether GIs can always be used to filter out unrelated sporadic cases from outbreak investigations.

5.2. Methods

For this analysis, I used a collection of 49 *L. monocytogenes* genome sequences shared with our lab from Public Health Canada (NML). These isolates represent various outbreaks, as classified by the NML, from across the country spanning 30 years. Table 5.1 highlights additional details about the outbreak isolates, including results from traditional laboratory methods to characterize and differentiate *L. monocytogenes* strains. All genomes used in this analysis were completely closed and two were previously published (08-5578 accession number NC_013766.1; 08-5923 accession number NC_013768.1). This collection of genomes allows for the examination of GI changes between various outbreaks (named outbreak 1-9), as well as a detailed comparison of the dynamics of GIs within a single large outbreak (outbreak 8). A set of sporadic isolates from human listeriosis cases that were not associated with any particular outbreak can also be used for comparison. IslandViewer 2 was used for GI prediction on all genomes (Dhillon et al., 2013) since IslandViewer 3 was still under development at the time. Nucleotide sequences for all GI predictions from the three methods were downloaded and used to perform comparative analyses between isolates. For each isolate, every predicted GI was used as input for a nucleotide BLAST search against every other isolate in the collection

for presence of that particular GI. Hits were required to have more than 80% sequence identity over 80% of the length of the GI with an e-value less than 1e-04. Cases where GIs overlapped partially were also marked. This BLAST analysis was used to generate a table of GI presence/absence that was used for visualization of GI clustering in R using the function “heatmap.2” within the “gplots” package.

Table 5.1 Select information about *L. monoctyogenes* outbreaks represented in genome sequencing collection

Outbreak	Number of isolates sequenced	Year	Location	Serotype	MLST	Clonal Complex
1	4	1981	Nova Scotia	4b	1	1
2	2	1996	Ontario	1/2b	5	5
3	2	2000	Manitoba	1/2a	7	7
4	2	2002	Quebec	1/2a	37	37
5	3	2002	British Columbia	4b	1	1
6	2	2002	British Columbia	4b	388	N/A
7	2	2008	Quebec	1/2a	394	415
8	12	2008	Ontario and Saskatchewan	1/2a	Multiple*	8
9	4	2010	Ontario	1/2a	120	8
Sporadic cases	16	1988-2011	N/A	1/2a, 3a	Multiple**	8

* Includes 120 and 292
**Includes 8, 120, 292, and 387

5.3. Results and Discussion

5.3.1. Types of GI predictions across outbreak isolates

In total, 33 predicted GIs were compared for presence/absence in all *L. monocytogenes* isolates used in this study, resulting in the identification of only 2 unique GIs and 31 GIs that were shared between a number of isolates. In light of the fact that boundary predictions for GIs sometimes require manual curation as they are not precise, some of the shared GI predictions were either predicted to be larger or smaller in between various isolates. Also, GIs clustered very closely could actually be a single larger GI.

These cases often require manual refinement of boundaries, but because this was such a large dataset across many genomes, I decided to use the boundaries as delimited by IslandViewer for consistency. All predicted GIs and a description of the gene products encoded by genes within these GIs can be found in Table 5.2. In addition to the many prophage structural genes, this includes multiple multidrug resistance genes, arsenic and other heavy-metal resistance genes, transport genes, transposases, recombinases, and hypothetical proteins.

Table 5.2 Description of GIs found in *L. monocytogenes* outbreak isolates

GI ID	Size (kb)*	Available gene product annotations of genes
GI_1	13,889	Hydantoinase/oxoprolinase, peptidoglycan linked protein, transposase, hypothetical proteins
GI_2	17,380	Arsenical resistance protein, putative antibiotic transport system ATP-binding protein, antibiotic transport permease protein, putative transcriptional regulator, heavy metal translocating P-type ATPase, hypothetical proteins
GI_3	4,303	Hypothetical proteins
GI_4	13,183	Membrane associated lipoprotein, transposase, hypothetical proteins
GI_5-	30,267	Integrase, transcriptional regulator, sugar-phosphate nucleotidyltransferase, single-strand DNA-binding protein, RNA polymerase sigma-70 factor, ECF subfamily, site-specific recombinase, phage terminase small subunit, terminase large subunit, portal protein, protease, phage capsid protein, phage protein, major tail protein, phage protein, minor structural protein GP75, phage minor structural protein, protein GP23, hypothetical proteins
GI_6	12,479	Hypothetical proteins
GI_7	16,524	Ribonuclease PH, nucleoside-triphosphatase, integrase, GP35, GP37, putative methyltransferase, putative phage integrase, putative phage protein, phage DNA/RNA helicase, hypothetical phage protein, hypothetical proteins
GI_8	24,693	Phage transcriptional regulator, single-stranded DNA-binding protein, phage protein, anti-repressor protein, transcriptional regulator, ATP-dependent DNA helicase RecG, integrase, competence protein K, phosphotransferase system (PTK) fructose-specific enzyme IIABC component, fructose-1-phosphate kinase, hypothetical phage proteins, hypothetical proteins
GI_9	11,149	Recombination protein RecT, anti-repressor protein, helix-turn-helix protein, integrase, competence protein K, phosphotransferase system (PTS) fructose-specific enzyme IIABC component, fructose-1-phosphate kinase, hypothetical phage protein, hypothetical protein
GI_10	6,316	Hypothetical proteins
GI_11	4,114	Hypothetical proteins
GI_12	6,953	4'-phosphopantetheinyl transferase, fatty-acyl-coA synthase, hypothetical protein

GI ID	Size (kb)*	Available gene product annotations of genes
GI_13	17,920	Transposase, putative secretive protein, 23S rRNA methyltransferase, heme-dregading monooxygenase, hypothetical protein
GI_14	5,804	Single-strand DNA binding protein, bacteriophage protein, phage transcriptional regulator, phage terminase, hypothetical phage protein, hypothetical protein
GI_15	13,237	Type I restriction enzyme M protein, type I restriction enzyme specificity protein, type I restriction enzyme, R subunit, restriction system protein, helicase, hypothetical proteins
GI_16	11,158	Mutt/NUDIX family protein, alkylphosphonate utilization operon protein PhnA, transcriptional regulator, β-glucosidase, lactose/cellobiose family IIC component protein, cellobiose-specific IIB component protein, RNA-directed DNA polymerase, DeoR family transcriptional regulator, hypothetical protein
GI_17	6,371	Hypothetical membrane protein, hypothetical protein
GI_18	19,647	Bifunctional GMP synthase/glutamine aminotransferase protein, ATPase, integrase, putative regulatory protein, glyoxalase family protein, acetyltransferase, hypothetical proteins
GI_19	14,895	DNA recombinase, recombinase, transcriptional regulator, multidrug resistance protein (SMR family), type IV secretion system, Sel 1 repeat protein, hypothetical protein
GI_20	11,300	Type IV secretory protein, hypothetical protein
GI_21	12,479	Pilus assembly protein CpaF, pilus assembly protein CpaB, late competence protein C, SpoVG family protein, stage V sporulation protein T, DNA adenine methylase, hypothetical protein
GI_22	9,766	Phage major capsid protein, phage minor capsid protein, phage terminase, phage transcriptional regulator, single-strand DNA-binding protein, sugar-phosphate nucelotidyltransferase, hypothetical phage protein, hypothetical protein
GI_23	9,983	Xylose repressor, methionyl-tRNA synthetase, sugar transport proteins, cell wall surface anchor family protein, transposase OrfA, transposase OrfB, internalin protein, hypothetical protein
GI_24	4,218	Phage major tail protein, Phi13 family, uncharacterized phage protein, phage major capsid protein, HK97 family, peptidase S14, ClpP, hypothetical proteins
GI_25	8,520	L-alanoyl-D-glutamate peptidase, phage holin, phage minor structural protein, phage tail protein, phage tail tape measure protein, TP901 family, hypothetical protein
GI_26	8,360	DNA recombinase, recombinase, transcriptional regulator, multidrug resistance protein, hypothetical protein
GI_27	8,530	Putative phage integrase, putative methyltransferase, DNA replication protein, transcriptional repressor of PBSX genes, hypothetical proteins

GI ID	Size (kb)*	Available gene product annotations of genes
GI_28	12.055	Phage protein, anti-repressor protein, transcriptional regulator, ATP-dependent DNA helicase RecG, integrase, competence protein K, hypothetical phage protein, hypothetical protein
GI_29	12,335	DNA methylase, S-adenosylmethionine synthetase, HNH endonuclease, SNF2-related protein, prophage antirepressor, hypothetical proteins
GI_30	32,894	Phosphotransferase enzyme family, K04763 integrase/recombinase XerD, AntA/AntB antirepressor domain protein, transcriptional regulator, Cro/CI family, K01159 crossover junction endodeoxyribonuclease ruvc, DNA polymerase I-3'-5' exonuclease and polymerase domains, DNA primase, K0234 replicative DNA helicase, DNA replication protein, transcriptional repressor of PBSX genes, hypothetical proteins
GI_31	5,315	NAD dependent epimerase/dehydratase, replication-associated protein repB, hypothetical protein
GI_32	7,775	Putative phage protein, conserved phage-related protein, phage protein, hypothetical proteins
GI_33	6,893	Peptidoglycan binding protein, hypothetical proteins

* Size of largest GI prediction listed here (size may vary between different isolates as boundary predictions are not precise)

5.3.2. Geographically and temporally distinct outbreaks harbour unique sets of GIs

Clustering of isolates using BLAST to detect the presence/absence of the 33 GIs in every isolate is shown in Figure 5.1. The overall trends reveal that there is a set of GIs that is shared by all the isolates, but there were specific differences between every outbreak, with the exception of outbreaks 8 and 9. Isolates within each outbreak shared identical sets of GIs that were unique to the outbreak and thus resulted in the formation of distinct outbreaks clusters. This could be due to many factors, but geographical and temporal separation could be key players in the type of genomic material that is available for acquisition in the environment. For instance, certain phage may be present in the environment or a particular food processing facility in British Columbia that aren't seen in Manitoba. This is likely the reason that outbreaks 8 and 9, which mainly occurred in Ontario between 2008 and 2010, share a lot of similarities and are clustered together because there may have been little geographic and temporal separation. Additional information was not available for investigating whether these outbreaks began from the exact same source location, which would be the most plausible explanation for the high similarity in GIs. Previously, a study of GIs in the fish pathogen *Aeromonas salmonicida*

subsp. *salmonicida* also revealed a geographical link to different forms of the *AsaGEI1* (a GI) that can be used to determine its origin (Emond-Rheault et al., 2015) and supports the observed geographic linkage of GIs in *L. monocytogenes* outbreak isolates.

On the contrary, there are also cases such as outbreaks 5 and 6, that occurred mainly in British Columbia in 2002, with limited geographic and temporal separation, that also have very distinct sets of GIs. It is difficult to imagine the circumstances of these differences without additional information about the outbreaks, however, I hypothesize that these outbreaks started from different locations or food processing facilities within British Columbia. It is also equally plausible that two different locations within a single food processing facility could have different micro-environments in which various genetic material is isolated and could only be spread to the *L. monocytogenes* within that micro-environment, for example, within a meat slicing machine. Altogether, it is evident from this GI clustering that GI content in outbreak isolates of *L. monocytogenes* generally varies between outbreaks and is likely driven by geographic and temporal separation of available genetic material in the environment.

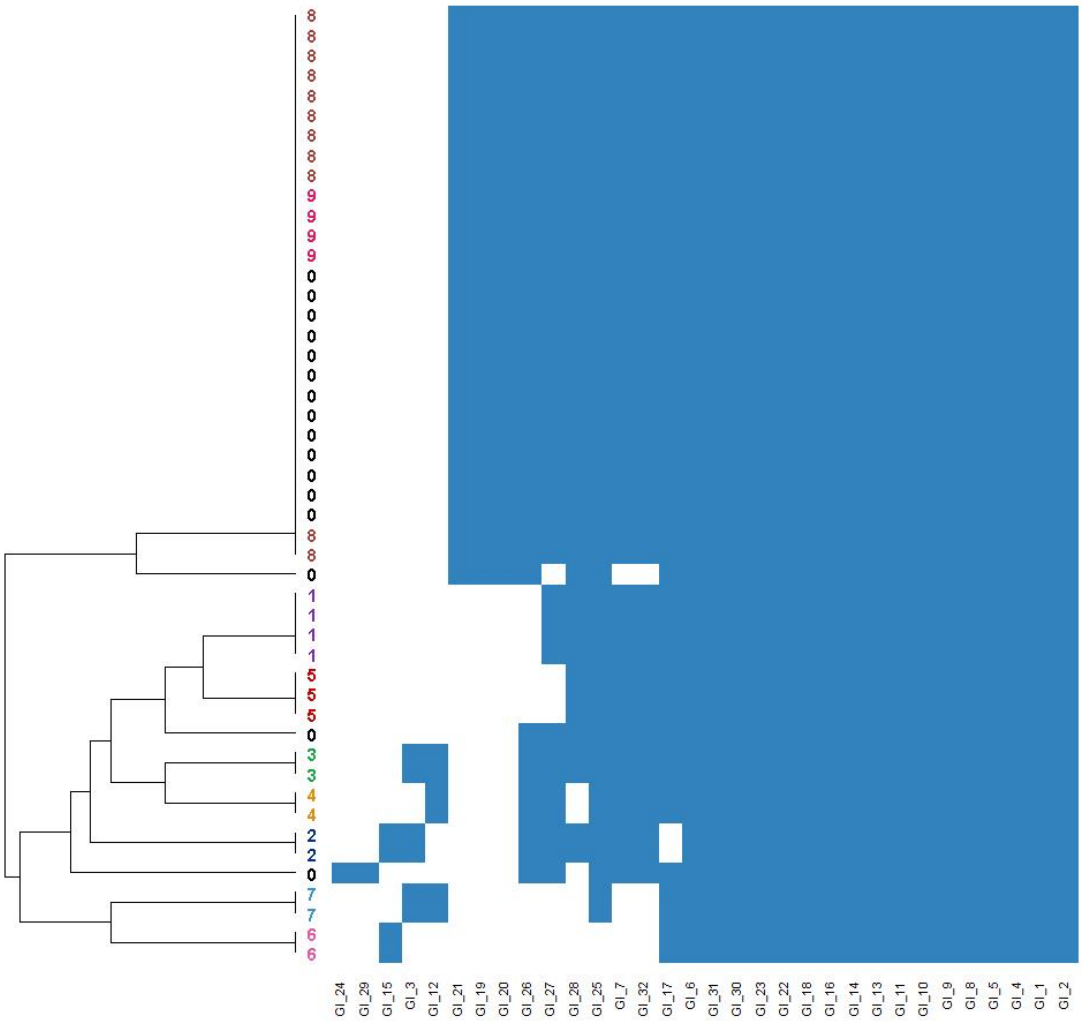

Figure 5.1 Clustering of predicted GIs in *L. monocytogenes* isolates
Each GI prediction was used as input for a BLAST search to confirm the presence or absence of the GI in every other isolate in the collection. Blue indicates the presence of the specified GI, while white means the GI is absent. Isolate names have been removed for privacy reasons and have been replaced by the designated outbreak number, where 0 indicates sporadic cases not associated with a particular outbreak.

Another important aspect of GI dynamics to consider is within a single outbreak. Outbreak 8 was the largest outbreak in our dataset that was found to be associated with ready-to-eat meat products in Canada in 2008, and resulted in 22 deaths and more than 50 cases of illness (Gilmour et al., 2010). Sequences for 12 isolates from this outbreak were used, plus a handful of sporadic cases that occurred around the same time (information about location was unavailable), to study the dynamics and stability of GIs

throughout this outbreak. From Figure 5.1, it is clear that outbreak 8 isolates all share an identical set of GIs. This outbreak doesn't have any unqiue GIs as it shares GIs with outbreak 9, however, together, these two outbreaks uniquely harbour GI_19, GI_20, and GI_21. 08-5578 and 08-5923 were previously reported to harbour a novel 49.8 kb GI (named Listeria genomic island 1, or LGI1) that encoded putative translocation, resistance and regulatory elements (Gilmour et al., 2010), and this roughly corresponds to the GI_19, GI_20, and GI_21 predictions from IslandViewer. What's more, outbreak 9 isolates and a large set of sporadic isolates also share all of these same GIs. This could indicate these outbreaks arose from very similar sources. Although, this figure is only presenting the presence/absence of the GIs and not the variation within each, upon further investigation it was found that this region is 99-100% identical across these isolates and no further clustering was possible based on this unique region to separate these outbreaks using GI content.

5.4. Conclusions

This chapter focused on using IslandViewer for GI prediction across a collection of isolates from food-borne outbreaks of *L. monocytogenes* from regions across Canada to highlight the importance of performing such analyses for outbreak investigations and to gain insight into the evolution of GIs from real outbreaks. First, using this dataset I was able to show that the presence/absence of GIs alone clusters geographically and temporally related outbreak isolates together. For those outbreaks that are not from disparate geographic regions and occur during relatively similar time periods, there is very little variation in GIs. This chapter highlights that horizontal gene transfer, which is an important source of variation for very closely related strains of *L. monocytogenes*, results in enough variation in GI content to be able to detect differences between outbreaks that are separated geographically and temporally, which would also filter out sporadic cases from different sources that may confound outbreak investigations.

However, this analysis is limited to one particular species and it is not known whether such trends exist for other infectious disease pathogens. For example, a *M. tuberculosis* outbreak in British Columbia did not exhibit any changes in GI content between isolates (Gardy et al., 2011) and may reflect differences in this pathogen's

exposure to, or interaction with, such mobile elements. Thus, additional evaluation of GI content across multiple outbreaks of distinct pathogens can further improve our understanding of the spread of GIs in the context of infectious disease. As more genomes become available for performing more robust analyses, the clustering of GIs could be compared against traditional MLST and PFGE typing data, or even SNV data to further exemplify the importance of GI analysis. The integration of IslandViewer into the IRIDA platform will be valuable in performing such additional evaluations using larger datasets across a variety of important pathogenic species.

The application of IslandViewer for this analysis also allowed the identification of issues concerning the development of better tools for performing comparative analyses of GI predictions across many more genomes. Manual comparisons of GI predictions using BLAST searches as performed in this chapter are tedious and time consuming. It is also not realistic to perform manual analysis for organisms that may have many more GIs, like Salmonella, for example. Alternative approaches using more recently developed algorithms for rapid whole genome alignments, such as Mugsy (Angiuoli and Salzberg, 2011) or Harvest (Treangen et al., 2014) may enhance comparisons of GI predictions across many more microbial genomes. Thus, the work presented in this chapter shows the dynamics of GI spread across outbreaks of *L. monocytogenes* and forms a basic framework for informing tools that are currently being developed for future functions of comparative analyses of GI predictions within IslandViewer across tens to hundreds, or even thousands, of genomes.

Chapter 6.

Genomic analysis of *Listeria monocytogenes* adapted to growth in cold environments

Patricia Hingston from Kevin Allen's Lab at the University of British Columbia conducted all phenotypic assays of growth described in section 6.2.1. Genomes were sequenced by Genome Quebec. I performed all other analyses presented in this chapter.

6.1. Introduction

As described in the previous chapter, *L. monocytogenes* is an important food-borne pathogen linked to multiple outbreaks (Swaminathan and Gerner-Smidt, 2007). Most concerning is that *L. monocytogenes* is a psychrotroph, with optimal growh at temperatures above 20°C but is able to grow at low temperatures (Walker et al., 1990) and can replicate to harmful levels on cold ready-to-eat foods, such as deli meats, within the shelf life (Nufer et al., 2007). Generally, low levels of *L. monocytogenes* (less than 100 CFU/g) pose very little risk of causing listeriosis (Chen et al., 2003), and are considered acceptable within the shelf life of ready-to-eat foods (Health Canada, 2011). However, the levels of tolerance to various stresses in food production environments is widely distributed among strains (Arguedas-Villa et al., 2014; Nufer et al., 2007; Van Der Veen et al., 2008). As such, growth studies on one or a collection of strains present on foods may not accurately represent the growth potential of the more tolerant strains that could be present. Therefore, better criteria and biomarkers need to be identified for assessing stress tolerance, especially to cold temperatures, and virulence potential of *L. monocytogenes* in food processing facilities.

In general, psychrotrophs and psychrophiles can harbour adaptive changes in lipids and proteins that allow them to maintain fluidity in cell membranes and to increase stability of enzymes performing metabolic reactions (Russell, 1990). For example, mutations that increase the hydrophobicity of interior regions of a protein can provide greater stability and may be favoured in psychrophiles and psychrotrophs (Russell, 1990).

Previous studies in *L. monocytogenes* specifically have found links between stress tolerance and particular genetic lineages, serotypes, or isolation sources, but they also find that the tolerance phenotype is not consistent as there are a number of strains with intermediate or sensitive phenotypes within these lineages, serotypes, or sources (Barbosa et al., 1994; Bergholz et al., 2010; Durack et al., 2013; Junttila et al., 1988; Ribeiro et al., 2014). Furthermore, cold tolerant isolates have been shown to share MLST and PFGE patterns with cold sensitive isolates, indicating smaller genetic changes such as SNVs could be responsible for these differences (Kovacevic et al., 2013). A study by Arguedas-Villa *et al.* in 2014 did identify five identical amino acid substitutions in the *sigL* protein in strains adapted to growth in cold from a collection of 62 isolates. Other SNV analyses have found SNVs causing premature stop codons within the *inlA* gene are in a higher percentage of intermediate cold adapting isolates (54%) versus cold tolerant (13%) and cold sensitive (20%). The *inlA* gene is an important VF in Listeria and truncation of this gene results in attenuated invasion of host cells (Nightingale et al., 2005), thus the cold tolerant isolates could also be more virulent than intermediate growers (Kovacevic et al., 2013).

To better understand the cold-growth adaptation in *L. monocytogenes*, 166 different isolates of *L. monocytogenes* collected from across Canada (namely from Alberta, British Columbia, and Nova Scotia) and Switzerland from food, food processing environments, and human cases were evaluated. The isolates were phenotypically assayed and sequenced in collaboration with Patricia Hingston and Jessica Chen from Dr. Kevin Allen's group at the University of British Columbia. Using the MLST data alone showed all 166 strains were high risk strains even though there were observed phenotypic differences in cold growth, so there is a need for the identification of other genetic markers that could be used as targets for profiling the risk associated with different strains of *L. monocytogenes*. I tested genetic variants, specifically SNVs, for significant associations with cold tolerance to determine potential mutations contributing to this phenotype, also known as a genome-wide association study (GWAS). In addition to this, I identified larger GI differences between the cold tolerant (CT) and cold sensitive (CS) groups in case cold-adaptation genes are shared via mobile elements.

GWAS have been very commonly applied to the study of human genetic diseases and disorders previously. The National Human Genome Research Institute (NHGRI) – European Bioinformatics Institute (EBI) GWAS Catalog (http://www.ebi.ac.uk/gwas/) reports 2,423 studies have been conducted and 16,617 unique associations have been identified as of April 17, 2016. The first microbial GWAS studies were reported in 2013 (Dutilh et al., 2013; Farhat et al., 2013; Sheppard et al., 2013), and have since shed light on the important differences in genome evolution that may affect such analyses. Bacteria have strong asexual population structure, have strong linkage disequilibrium, and phenotypes are generally the result of strong positive selection rather than genetic drift in human diseases and these factors should all be considered when performing GWAS in microbial genomes (Chen and Shapiro, 2015; Falush and Bowden, 2006; Read and Massey, 2014). In this study, the collection of isolates of *L. monocytogenes* represent geographically and temporally distinct populations that will hopefully minimize the identification of any false positive SNVs in association with the various tested phenotypes. In all, this chapter presents the first large-scale association study to identify genetic biomarkers related to cold tolerance in *L. monocytogenes*.

6.2. Methods

6.2.1. Cold growth assay

This assay was adapted from the procedure described by (Arguedas-Villa et al., 2010). All isolates of *L. monocytogenes* were grown in brain-heart infusion broth at 37°C overnight (16 hours) with shaking. Solutions were then standardized and serially diluted to 10^3 colony forming units per millilitre and incubated at either 37°C or 4°C. Optical density readings at a wavelength of 600nm were taken every 30 minutes for samples at 37°C or at days 0, 1, 2, 3, 4, 7, and then biweekly for samples at 4°C until stationary phase of growth was observed. Samples at 4°C were also plated on peptone saline (PS, 0.01% peptone, 0.85% NaCl) on tryptic soy agar (BD, Fisher Scientific) + 6% yeast extract (BD, Fisher Scientific) and were enumerated after growth at 37°C for 24 hours. Two biological replicates and two technical replicates were completed for each isolate and the growth curves were modeled using a four parameter logistic model that recorded the lag phase

duration, maximum growth rate (μ_{max}), and maximum viable cell density (N_{max}) (Dalgaard and Koutsoumanis, 2001). The μ_{max} (measured as (log CFU/ml)/hour) was standardized by dividing by the median value for the experimental run to account for variation between runs. These standardized values were averaged across replicates and isolates were classified as cold tolerant (CT) or cold sensitive (CS) if their average μ_{max} was greater or less than the median ±1 standard deviation, respectively. All other isolates within 1 standard deviation of the median were designated intermediate (INT) growers.

6.2.2. Genome analysis

Sequencing and quality control

Genomes were sequencing using the Illumina HiSeq platform by GenomeQuebec, which generated 100 base pair paired-end reads. The average insert size was 330 base pairs. The number of reads generated per isolate ranged between 5.2-16.8 million. As part of quality control, Trimmomatic version 0.25 (Bolger et al., 2014) removed low quality reads with default parameters except "-phred33 SLIDINGWINDOW:5:20 LEADING:20". Cutadapt version 1.5 (Martin, 2011) removed adapter sequences from reads using default parameters and specifying the Illumina TruSeq adapter sequences ("AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" for R1 and "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" for R2). FastQC version 0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) generated reports and visualizations of various quality control measures after each step to verify that each isolate's library of sequences was of high quality. Upon completion of quality control, each library had between 4.9-16.5 million high quality reads remaining.

Genome assembly

All 166 genomes were assembled *de-novo* using SPAdes (Bankevich et al., 2012) with the careful option, and contigs with less than 10x coverage or length less than 200 base pairs were filtered out. The remaining contigs were annotated using Prokka version 1.8 (Seemann, 2014), specifying that the genomes are Gram positive. A reference-based approach was also employed using the *L. monocytogenes* EGD-e reference genome (accession number NC_003210.1). This genome was chosen as a reference because of

two main reasons. For one thing, the 166 genomes represent different serovars and so it was not possible to find a reference genome that was closest to every isolate. Secondly, *L. monocytogenes* EGD-e is more extensively studied and better annotated from the *L. monocytogenes* reference genomes available to date and would be most useful for interpretation of variants. Raw reads from all genomes were mapped against the reference using SMALT version 0.7.6 (http://www.sanger.ac.uk/science/tools/smalt-0) with default parameters except "-i 330", specifying that the average insert size is 330 base pairs. SMALT has previously been shown to reduce the number of errors in assembling *Listeria* genomes that are not nearly identical to the reference (Pightling et al., 2014).

Mobile and accessory genome analysis

For GI detection, *de-novo* assembled genomes annotated with Prokka were submitted in batches through the local command line interface for the custom genomes pipeline of IslandViewer 3 (Dhillon et al., 2015). Upon completion, the integrated GI predictions from IslandPath-DIMOB, SIGI-HMM and IslandPick were downloaded from the web server and further analyzed for overlap. More specifically, all genes predicted to be on GIs in CT isolates were used as input for a protein BLAST search against all other isolates to identify any GIs unique to the CT isolates. Roary version 3.0.2 (Page et al., 2015) was used to identify pan genomes, defined as the union of all genes shared by genomes of interest (Medini et al., 2005; Vernikos et al., 2015), from the *de-novo* assembled genomes. This tool also enables the identification of core genes shared by every genome and accessory genes that are found in subsets of genomes. Roary was used to identify genes unique to various phenotypic groups using the command "query_pan_genome –a difference" and defining two input sets based on phenotypic measurements, for example, CT isolates were compared against CS isolates.

Phylogenetic reconstruction based on core genome SNVs

The Harvest suite of tools is designed to quickly analyze thousands of microbial genomes by performing core genome alignment, variant calling, recombination detection and building phylogenetic trees (Treangen et al., 2014). Parsnp, a tool within Harvest, was used to perform core genome alignment of all *de-novo* assembled genomes and the reference *L. monocytogenes* EGDe strain (NC_003210.1). This tool works best on assembled genomes rather than raw reads so nucleotide fasta files of all assembled

contigs were used as input. Core genome SNVs were detected based on the alignments and a filter removed SNVs clustered very closely together, which may be sequencing or alignment errors within repetitive regions of the genome (Reumers et al., 2012). For other microbial species, like *M. tuberculosis*, a filter to remove SNVs clustered within 50 base pairs works well (from presentation "SNP Calling & Outbreak Reconstruction in a Monomorphic Pathogen" by Gardy, J.L. at American Society for Microbiologists Conference Workshop in 2015), however, for *L. monocytogenes*, this value was too high and removed too many SNVs to differentiate between isolates. Thus, SNVs clustered within 20 base pairs were removed and the remaining high quality SNVs predicted from the core genome alignments were used to build maximum likelihood phylogenetic trees using RaxML version 8 (Stamatakis, 2014) on the CIPRES science gateway (Miller et al., 2010). Phenotypic data was also overlaid on top of phylogenetic trees for better visualization of trends using GraPHlAn (Asnicar et al., 2015).

SNV detection

Using the reference-based genome assemblies, Samtools version 1.2 (Li, 2011) was used to sort the aligned reads, remove any potential PCR duplicates, and call SNVs that were found against the reference genome. SNVs were further filtered to remove potential erroneous calls. The average coverage for each genome was roughly 500x (ranged between 300x and 1000x) and so any SNV calls with a read depth less than 50 were removed. Furthermore, since we know these genomes are haploid, any erroneous heterozygous SNV calls were also removed. This also removes mixed calls that may be the result of intra-host microevolution or subpopulations that may be present. An additional filter removed SNVs in repetitive regions that were identified by the index of repetitiveness (Haubold and Wiehe, 2006). Upon completion of the SNV filtering, remaining high-confidence SNVs were annotated using SNPEff version 4.1 (Cingolani et al., 2012) with the Listeria_monocytogenes_EGD_e_uid61583 annotation available within the program's database. In the end, for identification of potential SNVs contributing to differences in adaptation to cold, synonymous SNVs were also removed so that only non-synonymous variants and potential intergenic regulatory variants remained.

6.2.3. Statistical methods for identifying phenotype-genotype associations

SNPSift version 4.1 “CaseControl” command first enumerates the number of isolates from various case versus control groups and then runs basic p-value calculation using Fisher Exact test (as presented in “CC_ALL” by CaseControl) to identify SNVs that are statistically significantly associated with the case group isolates. For example, to identify the SNVs only found in CT *L. monocytogenes*, the tolerant isolates were included in the case group, and the isolates unable to grow in cold temperatures, CS, were the controls. This method is great for picking up SNVs with strong associations with case or control groups; however, certain associations may require very large sample sizes to be statistically significant, especially in cases where there is underlying genetic heterogeneity resulting in the same phenotype. Another disadvantage of this method is that each SNV is tested individually and thus it is not possible to identify interacting SNVs that may be contributing to the resulting phenotype.

Random Forests™ (Breiman, 2001), a machine learning method, can rank SNVs to identify those that are most important in classifying genomes for a given phenotype. A random forest is a collection of decision trees that are built using bootstrapping by selecting a random set of the features or variables that are being tested (with replacement). It is an important method to extract signals from complex biological datasets (Touw et al., 2013). Previous GWAS studies have successfully applied random forests and have been shown to perform better than the Fisher Exact test when multiple SNVs interact or there is genetic heterogeneity in acquiring a phenotype (Bulinski et al., 2011; Bureau et al., 2005; De Lobel et al., 2010; Lunetta et al., 2004; Schwender et al., 2004). The RandomForest™ version 4.6-10 library was used in R version 2.15.1 with a list of isolates, their cold-phenotype classification, and genotype at various SNV positions as input. This allows the method to run in a supervised mode, essentially as a classifier, which would rank SNVs in their ability to classify the cold phenotypes. The “randomForest” command was run with default parameters except “importance=TRUE, proximity=TRUE, ntree=5000”. The “ntree” variable should grow with the number of variables used to classify (Liaw and Wiener, 2002), thus I chose to set “ntree” relatively high.

6.3. Results and Discussion

6.3.1. Phenotypic variability of *L. monocytogenes isolates*

The doubling time of *L. monocytogenes* at 37°C in BHI broth is roughly between 45 to 60 minutes (Jones and D'Orazio, 2013). At 4°C, the raw μ_{max} values observed for this collection of *L. monocytogenes* isolates ranged between 0.017 and 0.052, which is equivalent to a doubling time of roughly 7.5 hours. Based on the standardized μ_{max} values, each isolate was assigned one of five designations: 1) very cold tolerant (VCT), 2) cold tolerant (CT), 3) intermediate (INT), 4) cold sensitive (CS), or 5) very cold sensitive (VCS). The difference in standardized μ_{max} between all of these designated groups was not clearly visibly distinct, but if INT growers are removed, as seen in Figure 6.1, the VCT and CT isolates together have higher standardized μ_{max} than the VCS and CS isolates, especially the VCS. For the remaining analyses, I will group the VCT and CT isolates together and consider them as the CT group, and the VCS and CS isolates as the CS group.

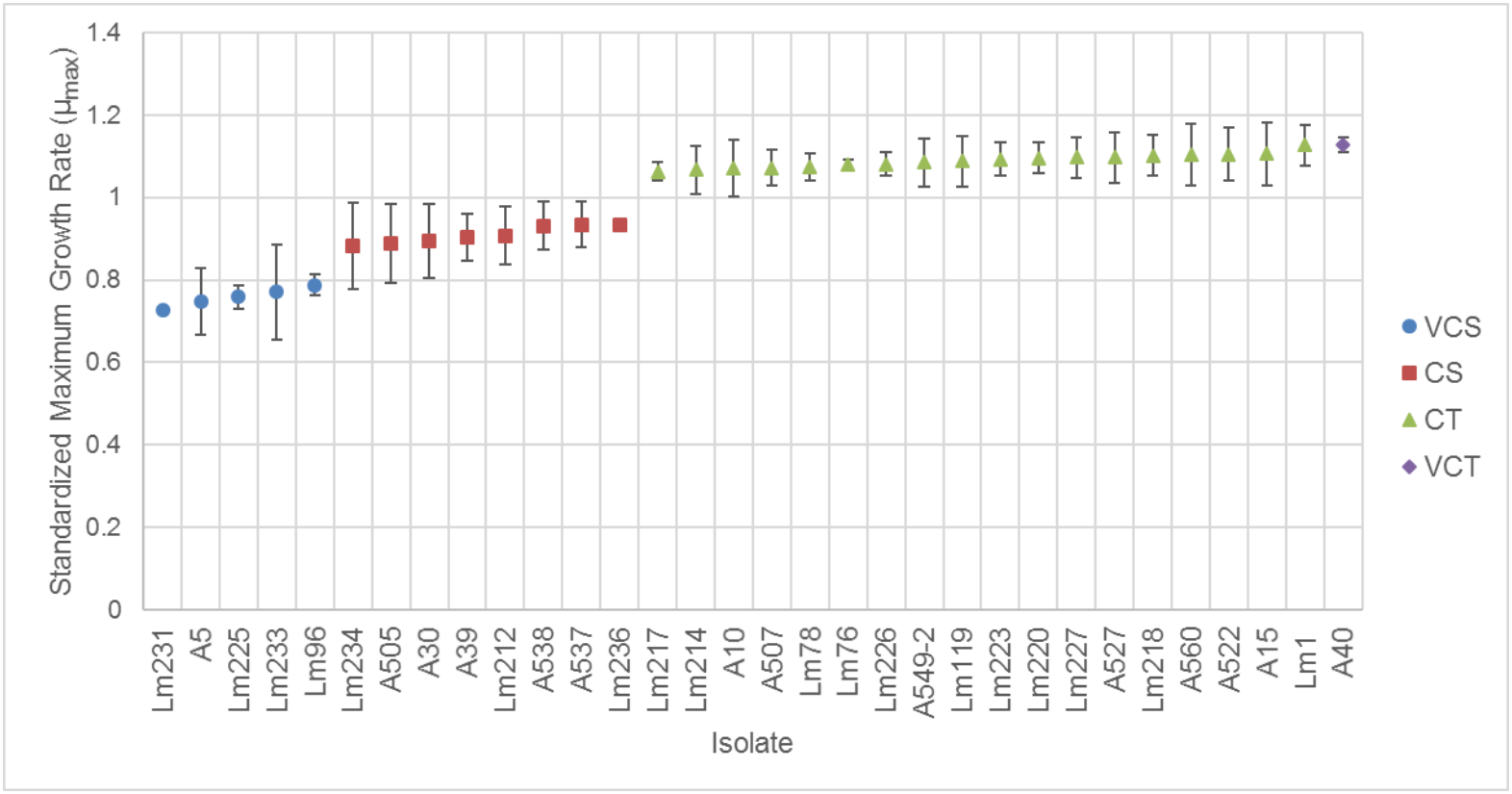

Figure 6.1 Standardized maximum growth rate (μ_{max}) of sensitive and tolerant isolates in cold growth assay at 4°C

6.3.2. Phylogenetic clustering of cold tolerant isolates

In order to better understand the relatedness and divergence between the various strains of *L. monocytogenes* in this collection and to determine whether phenotypically similar isolates are more closely related to each other, I built a number of phylogenetic trees based on SNVs found in the core genome. Figure 6.2 is the RaxML phylogenetic tree of all 166 isolates with the reference *L. monocytogenes* EGDe (NC_003210.1) as the root of the tree. Only 23% of the genome was identified as being core between all genomes, indicating that there is extensive variation between the isolates. This is likely because many different serovars are represented in this collection, so the strains were consequently separated based on serovar type for building better resolved trees. Using Figure 6.2 as a guide, serovars 1/2a, 1/2c, and 3a were grouped together, while serovars 1/2b, 4b, and 4c were placed in a separate group to generate trees with higher resolution using a larger dataset of high quality core genome SNVs. In doing so, the core genome of serovars 1/2a, 1/2c, and 3a was measured at 62% and 72% for serovars 1/2b, 4b, and 4c, including the reference genome in both cases.Figure 6.3 and Figure 6.4 represent the separated trees in radial view with cold growth phenotypic designations integrated on top to help identify clusters of phenotypically similar isolates. Cladogram views of these trees are available in Appendix B. It is certainly clear from these figures that CT and CS strains are not restricted to any one particular serovar or clade, indicating that this CT phenotype arose by convergent evolution possibly through multiple different mechanisms. However, there are also clusters on the trees, such as the serovar 1/2c isolates in Figure 6.3, that are very closely related but have highly variable cold growth phenotypes. These trends in the phylogenetic trees suggest overall that there may be multiple mechanisms utilized by *L. monocytogenes* to adapt to growth in cold temperatures, and the core genome SNVs do not reliably resolve these phenotypic differences. This phenotype may be the result of convergent evolution through the accumulation of beneficial mutations in different strains, but could also be acquired through horizontally transferred genes that confer an advantage in cold growth and will be discussed in further detail in the following sections.

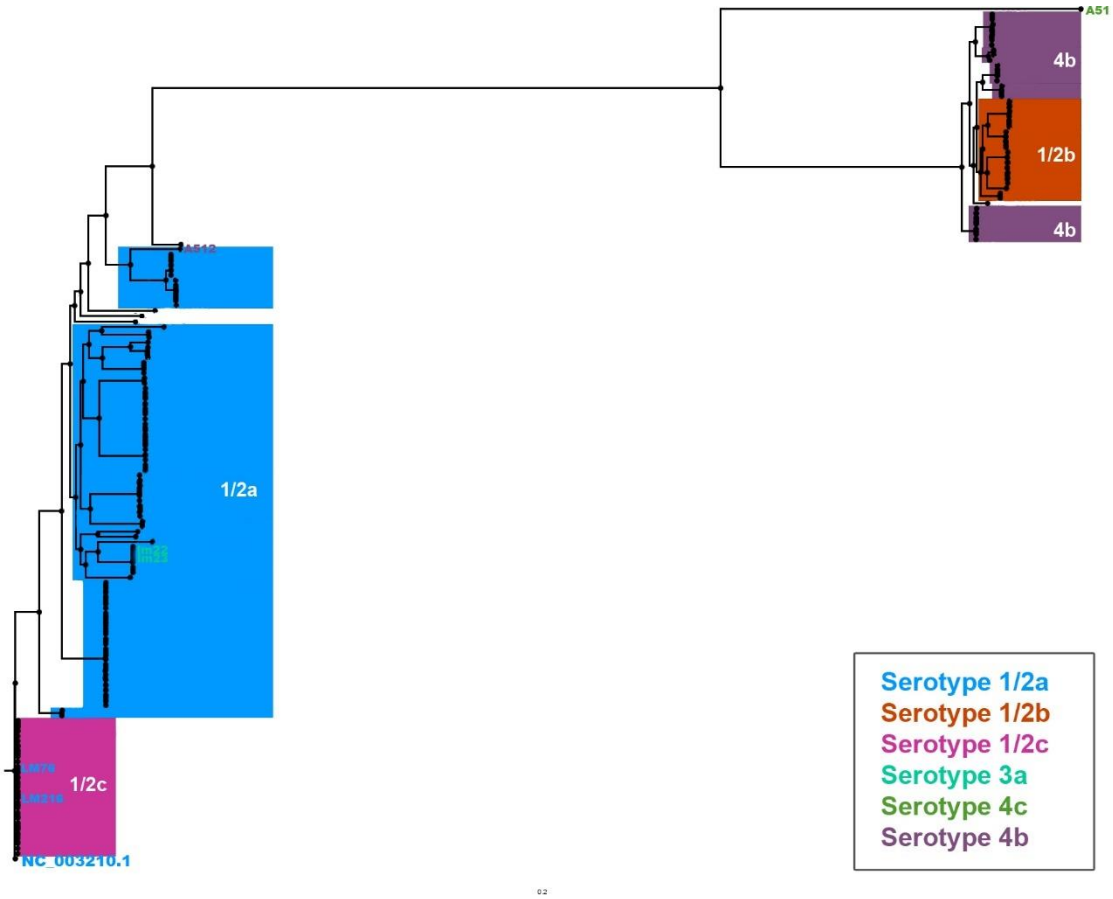

Figure 6.2 High-level overview of phylogenetic clustering of *L. monocytogenes* isolates in cold growth study collection.

Tree was generated using RaxML with core genome SNVs as input Isolates have been highlighted/coloured by serovar (see legend within figure). Serovars 1/2b and 4b and 4c have clearly diverged away from serovars 1/2a, 1/2c and 3a.

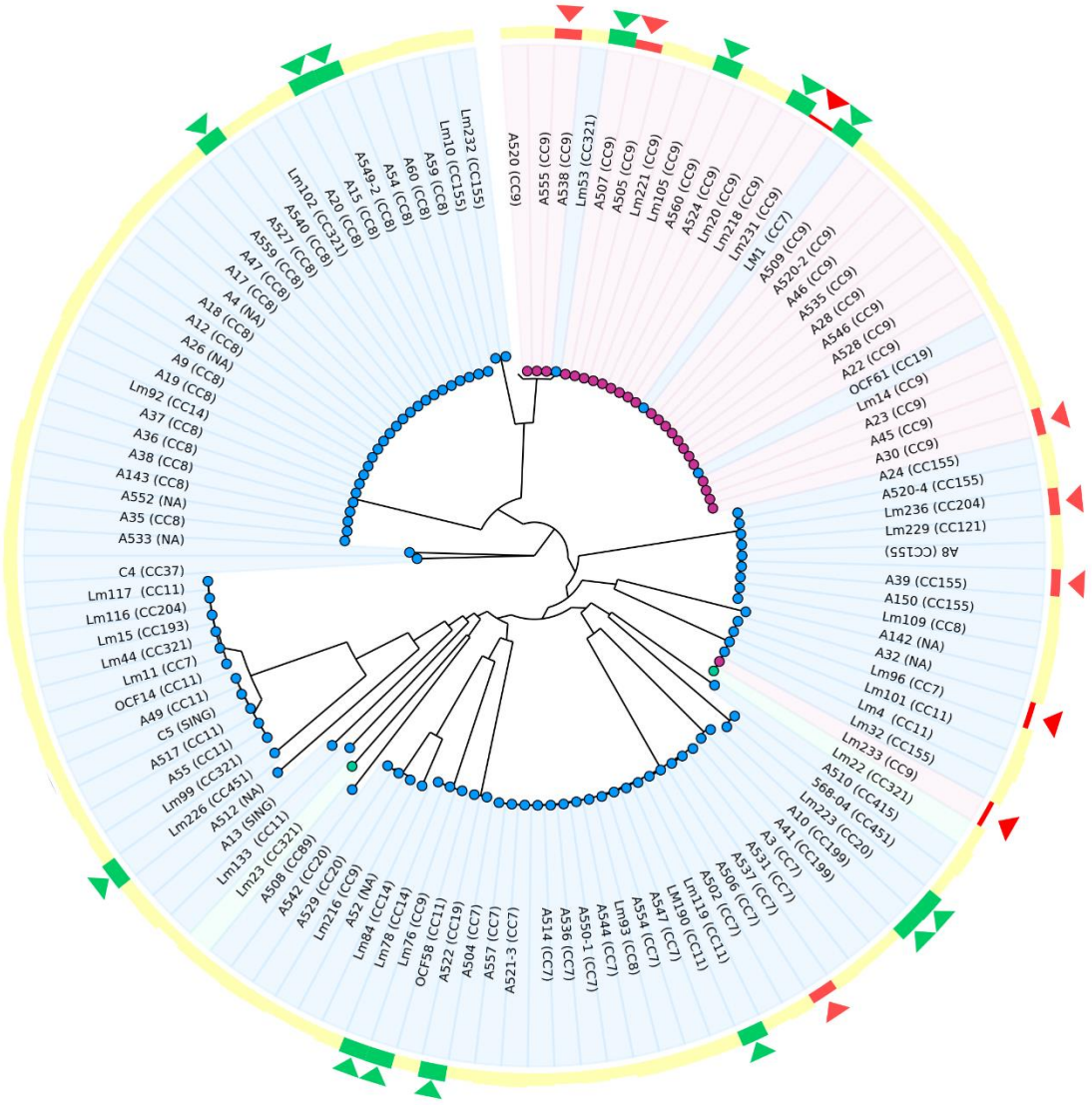

Figure 6.3 Phylogenetic tree of *L. monocytogenes* isolates of serovars 1/2a, 1/2c, and 3a using core genome SNVs calculated by Parsnp

Isolates on the tree are highlighted by serovar: 1/2a isolates are in blue, 1/2c isolates are in pink, and 3a isolates are in light green. The arrows in the outermost ring identify CS isolates in red and CT isolates in green. The histogram bars in the next ring in represent the standardized μ_{max} also coloured by designation: CT phenotype in green, CS phenotype in red and INT growers in yellow.

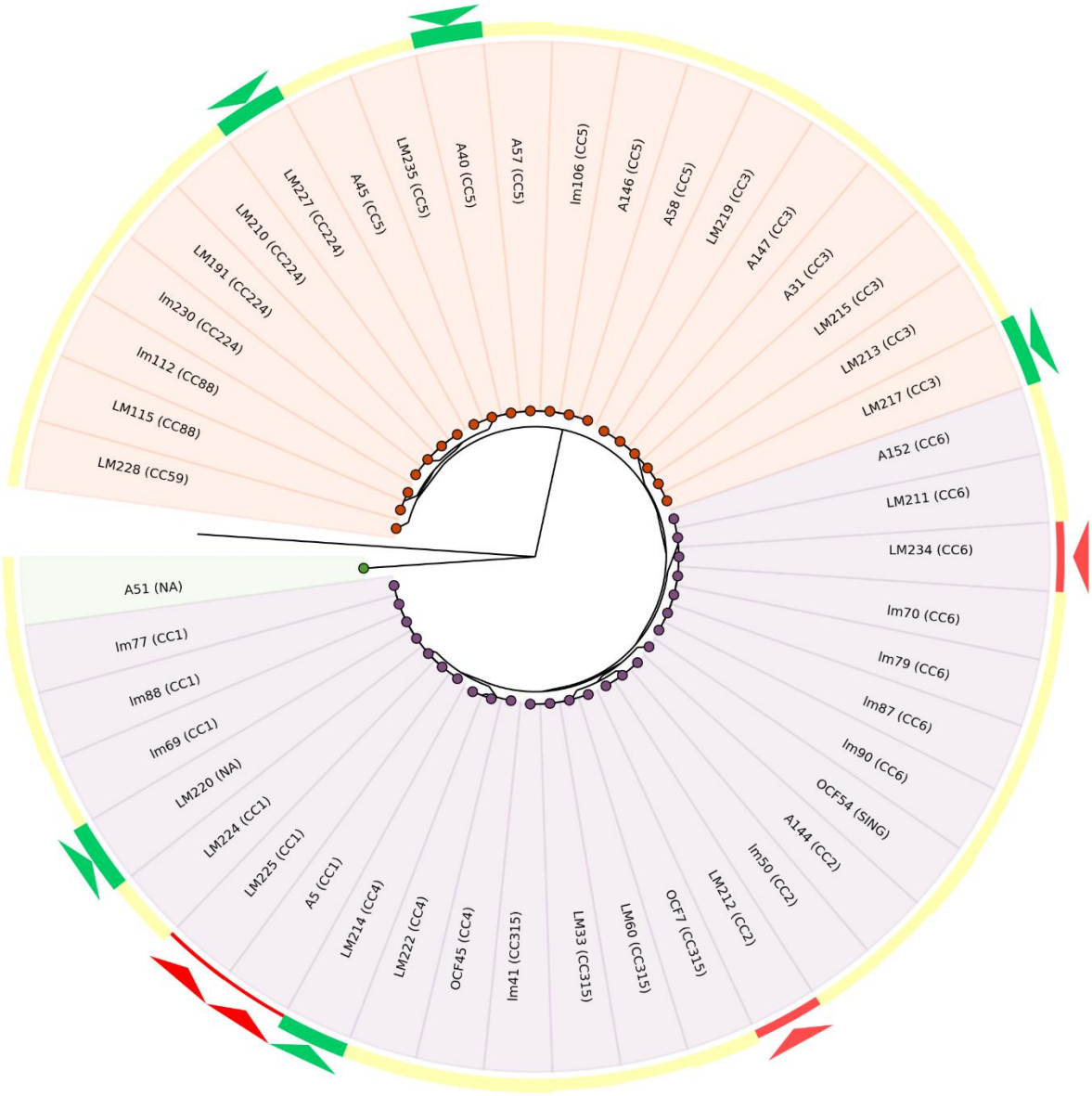

Figure 6.4 Phylogenetic tree of *L. monocytogenes* isolates of serovars 1/2b, 4c, and 4b using core genome SNVs calculated by Parsnp.

Isolates on the tree are highlighted by serovar: 1/2b isolates are in orange, 4c isolates are in light green, and 4b isolates are in purple. The arrows in the outer ring identify CS isolates in red and CT isolates in green. The histogram bars in the next ring in represent the standardized μ_{max} also coloured by designation: CT phenotype in green, CS phenotype in red and INT growers in yellow.

6.3.3. Role of mobile and accessory genes in cold growth

To identify genomic differences between CT and CS groups, first the isolates were compared at the gene level, to test whether CT isolates harbour any additional genes that

may be contributing to their ability to thrive in cold temperatures. Certainly the phylogenetic analysis suggests HGT could be playing a role in spreading cold-tolerance related genes as the phenotype is sporadically distributed in the phylogeny of these isolates. However, Roary analysis to compare core and accessory genomes between CT and CS isolates revealed no difference in gene content between these groups. Furthermore, IslandViewer analysis of this collection of *L. monocytogenes* isolates revealed the presence of many different GIs, including multiple different phage, however, there was no single GI shared by all CT isolates. This result reveals that the ability of some isolates of *L. monocytogenes* to grow in cold temperatures is likely not encoded on additional genes and therefore, there must be other genetic variation accounting for this difference in phenotypes.

6.3.4. SNVs associated with cold tolerance

To test the hypothesis that CT isolates harbour particular SNVs in key genes that are involved in tolerance to growth at cold temperatures, first SNPSift “CaseControl” command was used on the SNVs predicted from the reference-based method. This method did not identify a single non-synonymous SNV found in *all* CT isolates that was never seen in the CS isolates. Based on the previous phylogenetic analysis, this was not too surprising. We would expect to see serovar-specific changes or subgroups of SNVs that contribute to cold tolerance. What’s more, there was no non-synonymous SNV unique to a CT isolate that wasn’t also seen in any INT or CS isolates. Therefore, in Table 6.1, I have presented 110 non-synonymous SNVs found in subsets of CT isolates that are never seen in CS isolates but were seen in a number of INT growers. This table is sorted by p-value as calculated using a Fisher’s Exact test to indicate the significance of the association of the SNV with the CT isolates versus the CS isolates. To summarize, these SNVs affect 73 different genes, including a handful of VFs (specifically internalins), AMR genes (ABC transporter ATP-binding proteins), many hypothetical proteins and transcriptional regulators and their role in cold tolerance warrants further investigation. Below, I will discuss a few of the most significant SNVs associated with the CT and INT groups as found by this method and whether they could play a role in the cold growth phenotype.

Table 6.1 Non-synonymous SNVs found in a number of CT isolates but never in CS isolates.

SNV position	Gene	Locus tag	Protein accession	Reference Base	SNV base	Amino acid change	CT isolates	CS isolates	INT isolates	p-value of association
283379	inlG	lmo0262	NP_463793.1	T	A	Leu209Ile	8	0	40	2.94E-06
61689		lmo0057	NP_463590.1	G	A	Ala248Thr	9	0	51	9.32E-06
358406		lmo0331	NP_463861.1	T	C	Asn424Ser	9	0	52	9.32E-06
283800	inlG	lmo0262	NP_463793.1	C	T	Thr349Met	7	0	37	1.03E-05
353091		lmo0327	NP_463857.1	A	G	Thr545Ala	8	0	43	4.55E-05
1368871		lmo1341	NP_464866.1	T	C	Ile39Val	8	0	45	4.55E-05
1830590	pcrA	lmo1759	NP_465284.1	T	C	Ile530Val	8	0	49	4.55E-05
2049168		lmo1976	NP_465500.1	A	G	Tyr83His	8	0	51	4.55E-05
2896388		lmo2812	NP_466334.1	G	A	Ala50Thr	8	0	38	4.55E-05
2907552		lmo2821	NP_466343.1	C	A	Gln134Lys	8	0	44	4.55E-05
1114935		lmo1080	NP_464605.1	T	G	Ser393Ala	8	0	47	1.05E-04
392048		lmo0365	NP_463895.1	G	A	Val149Ile	9	0	50	1.20E-04
79682		lmo0075	NP_463608.1	G	A	Met212Ile	7	0	49	2.02E-04
81688		lmo0078	NP_463611.1	A	G	Lys10Glu	7	0	40	2.02E-04
82205		lmo0078	NP_463611.1	A	C	Glu182Ala	7	0	41	2.02E-04
82420		lmo0078	NP_463611.1	G	A	Val254Ile	7	0	41	2.02E-04
82543		lmo0078	NP_463611.1	G	A	Ala295Thr	7	0	40	2.02E-04
82544		lmo0078	NP_463611.1	C	A	Ala295Glu	7	0	40	2.02E-04
82552		lmo0078	NP_463611.1	A	C	Lys298Gln	7	0	40	2.02E-04
343913		lmo0319	NP_463849.1	T	G	Asn242His	7	0	46	2.02E-04
357845		lmo0331	NP_463861.1	G	A	Ser611Leu	7	0	36	2.02E-04
366980		lmo0334	NP_463864.1	A	G	Asp158Gly	7	0	44	2.02E-04

SNV position	Gene	Locus tag	Protein accession	Reference Base	SNV base	Amino acid change	CT isolates	CS isolates	INT isolates	p-value of association
471950		lmo0441	NP_463970.1	C	T	Ser70Asn	7	0	42	2.02E-04
898587		lmo0859	NP_464385.1	G	A	Glu439Lys	7	0	44	2.02E-04
1029777		lmo0998	NP_464523.1	C	T	Val15Ile	7	0	41	2.02E-04
1068037		lmo1037	NP_464562.1	C	T	Val164Ile	7	0	47	2.02E-04
1165608		lmo1131	NP_464656.1	A	G	Asn455Ser	7	0	33	2.02E-04
1338559		lmo1311	NP_464836.1	G	T	Thr130Asn	7	0	44	2.02E-04
1368706		lmo1341	NP_464866.1	C	G	Glu94Gln	7	0	34	2.02E-04
1368758		lmo1341	NP_464866.1	A	C	Asp76Glu	7	0	34	2.02E-04
1369513		lmo1343	NP_464868.1	C	T	Val62Ile	7	0	41	2.02E-04
1422055		lmo1394	NP_464919.1	G	T	Arg201Leu	7	0	37	2.02E-04
1445682		lmo1415	NP_464940.1	C	G	Ala164Gly	7	0	33	2.02E-04
1647458		lmo1603	NP_465128.1	G	A	Ala264Thr	7	0	49	2.02E-04
1793883		lmo1729	NP_465254.1	C	T	Met457Ile	7	0	42	2.02E-04
1793884		lmo1729	NP_465254.1	A	G	Met457Thr	7	0	42	2.02E-04
1793899		lmo1729	NP_465254.1	G	T	Ala452Glu	7	0	42	2.02E-04
2118023	murD	lmo2036	NP_465560.1	C	T	Ala217Thr	7	0	42	2.02E-04
2164953		lmo2086	NP_465610.1	T	C	Asn142Ser	7	0	44	2.02E-04
2164978		lmo2086	NP_465610.1	C	T	Glu134Lys	7	0	44	2.02E-04
2191830		lmo2111	NP_465635.1	T	C	Thr143Ala	7	0	42	2.02E-04
2191871		lmo2111	NP_465635.1	T	C	Asp129Gly	7	0	43	2.02E-04
2262629		lmo2178	NP_465702.1	T	A	Ile655Phe	7	0	45	2.02E-04
2262659		lmo2178	NP_465702.1	C	T	Asp645Asn	7	0	45	2.02E-04
2262660		lmo2178	NP_465702.1	A	C	Ile644Met	7	0	45	2.02E-04
2262739		lmo2178	NP_465702.1	G	C	Ala618Gly	7	0	45	2.02E-04

SNV position	Gene	Locus tag	Protein accession	Reference Base	SNV base	Amino acid change	CT isolates	CS isolates	INT isolates	p-value of association
2262740		lmo2178	NP_465702.1	C	T	Ala618Thr	7	0	45	2.02E-04
2312182		lmo2222	NP_465746.1	C	T	Ser172Asn	7	0	46	2.02E-04
2477751		lmo2404	NP_465927.1	A	T	Thr123Ser	7	0	40	2.02E-04
2883792		lmo2798	NP_466320.1	T	C	Glu96Gly	7	0	35	2.02E-04
2883807		lmo2798	NP_466320.1	T	C	Glu91Gly	7	0	35	2.02E-04
2915335		lmo2827	NP_466349.1	T	G	Asn103His	7	0	43	2.02E-04
804693		lmo0780	NP_464307.1	T	G	Ser104Arg	6	0	44	2.52E-04
357927		lmo0331	NP_463861.1	C	T	Gly584Arg	7	0	36	7.49E-04
2387069		lmo2303	NP_465827.1	G	A	Leu138Phe	6	0	27	7.82E-04
2387097		lmo2303	NP_465827.1	C	A or T	Arg128Ser or Arg128Arg	6	0	29	7.82E-04
141796		lmo0140	NP_463673.1	A	C	Asn60Thr	6	0	31	8.24E-04
355321		lmo0327	NP_463857.1	C	A	Pro1288Gln	6	0	28	8.24E-04
366877		lmo0334	NP_463864.1	T	A	Leu124Met	6	0	34	8.24E-04
395704		lmo0368	NP_463898.1	G	C	Val35Leu	6	0	29	8.24E-04
406104		lmo0382	NP_463912.1	G	T	Thr113Lys	6	0	38	8.24E-04
406116		lmo0382	NP_463912.1	G	A	Ala109Val	6	0	38	8.24E-04
446021		lmo0425	NP_463954.1	A	G or C	Arg325Gly or Arg325Arg	6	0	26	8.24E-04
466887		lmo0437	NP_463966.1	T	A	Val124Glu	6	0	20	8.24E-04
535867		lmo0500	NP_464028.1	C	G	Thr147Ser	6	0	43	8.24E-04
810520		lmo0785	NP_464312.1	G	C or A	Asp91Glu or Asp91Asp	6	0	39	8.24E-04
956485		lmo0919	NP_464445.1	A	C	Lys152Thr	6	0	24	8.24E-04
982390		lmo0947	NP_464472.1	G	T	Pro191Thr	6	0	45	8.24E-04
1019749		lmo0989	NP_464514.1	C	T	Leu69Phe	6	0	39	8.24E-04
1161015		lmo1126	NP_464651.1	A	G	Ile80Val	6	0	36	8.24E-04

SNV position	Gene	Locus tag	Protein accession	Reference Base	SNV base	Amino acid change	CT isolates	CS isolates	INT isolates	p-value of association
1164962		lmo1131	NP_464656.1	A	G	Ile240Val	6	0	40	8.24E-04
1173840		lmo1141	NP_464666.1	T	A or G	His66Gln or His66Gln	6	0	40	8.24E-04
1357531	truB	lmo1328	NP_464853.1	G	A	Gly255Ser	6	0	29	8.24E-04
1422082		lmo1394	NP_464919.1	C	G	Ala210Gly	6	0	27	8.24E-04
1424230	cinA	lmo1397	NP_464922.1	C	G	Thr110Ser	6	0	36	8.24E-04
1461641		lmo1430	NP_464955.1	T	G	Lys29Thr	6	0	36	8.24E-04
1486301		lmo1453	NP_464978.1	C	T	Val38Ile	6	0	36	8.24E-04
1530439		lmo1499	NP_465024.1	T	A	Tyr29Phe	6	0	33	8.24E-04
1754039		lmo1689	NP_465214.1	C	T	Asp302Asn	6	0	37	8.24E-04
1777821		lmo1716	NP_465241.1	T	A	Leu20His	6	0	37	8.24E-04
1830232	pcrA	lmo1759	NP_465284.1	T	C	Asn649Ser	6	0	35	8.24E-04
1973275	dinG	lmo1899	NP_465423.1	T	C	Asp563Gly	6	0	32	8.24E-04
2152513		lmo2073	NP_465597.1	C	T	Ala257Val	6	0	30	8.24E-04
2159277		lmo2080	NP_465604.1	C	A	Ala110Ser	6	0	37	8.24E-04
2164819		lmo2086	NP_465610.1	G	T or A	Leu187Ile or Leu187Phe	6	0	39	8.24E-04
2165064		lmo2086	NP_465610.1	T	C or A	Gln105Arg or Gln105Leu	6	0	45	8.24E-04
2165069		lmo2086	NP_465610.1	C	A or G	Glu103Asp or Glu103Asp	6	0	45	8.24E-04
2165074		lmo2086	NP_465610.1	C	T	Val102Ile	6	0	45	8.24E-04
2165092		lmo2086	NP_465610.1	T	C	Lys96Glu	6	0	45	8.24E-04
2165115		lmo2086	NP_465610.1	G	A	Pro88Leu	6	0	45	8.24E-04
2889115		lmo2804	NP_466326.1	A	G	Ile388Thr	6	0	35	8.24E-04
2896485		lmo2812	NP_466334.1	A	G or C	Lys82Arg or Lys82Thr	6	0	21	8.24E-04
2907385		lmo2821	NP_466343.1	C	T	Ala78Val	6	0	40	8.24E-04
2915833		lmo2828	NP_466350.1	T	C	Met89Val	6	0	33	8.24E-04

SNV position	Gene	Locus tag	Protein accession	Reference Base	SNV base	Amino acid change	CT isolates	CS isolates	INT isolates	p-value of association
2918580		lmo2832	NP_466354.1	G	A	Pro197Leu	6	0	32	8.24E-04
1214280		lmo1188	NP_464713.1	G	T	Met215Ile	7	0	47	9.42E-04
1313739		lmo1289	NP_464814.1	G	T	Ser29Ile	5	0	32	1.30E-03
889082		lmo0849	NP_464375.1	T	C	Thr141Ala	5	0	26	1.62E-03
778329		lmo0751	NP_464278.1	C	A	Ala23Glu	4	0	26	1.89E-03
171836		lmo0171	NP_463704.1	G	A	Arg776Lys	5	0	39	2.14E-03
1153758		lmo1116	NP_464641.1	T	C	Ile250Thr	4	0	29	2.50E-03
2394639		lmo2319	NP_465843.1	G	A	Ala52Val	4	0	22	3.05E-03
2394641		lmo2319	NP_465843.1	A	C	Asp51Glu	4	0	22	3.05E-03
1158936		lmo1123	NP_464648.1	T	C	Phe3Leu	4	0	28	3.55E-03
145915		lmo0147	NP_463680.1	A	C	Tyr16Ser	4	0	26	4.12E-03
2365371	lysA	lmo2278	NP_465802.1	C	T or A	Ala77Thr or Ala77Ser	2	0	22	1.96E-02
1142657		lmo1106	NP_464631.1	T	G or A	Leu631Phe or Leu631Phe	5	0	26	5.73E-02
2390439		lmo2312	NP_465836.1	C	G or A	Gly49Arg or Gly49Cys	2	0	5	6.97E-02
1142651		lmo1106	NP_464631.1	T	A or C	Lys633Asn or Lys633Lys	5	0	25	1.76E-01

inlG (NP_463793.1) has two of the most significant SNV positions in discriminating between CT and CS groups. The *inlG* protein is part of a group of proteins called the Internalins, specific to Listeria, that harbour leucine-rich repeats (LRRs) and play a role in attachment and invasion of host cells (Bierne et al., 2007). The first SNV at position 238379 results in a leucine to isoleucine substitution at position 209, which is not a drastic difference. Leucine and isoleucine generally can replace each other without largely affecting protein structure, even in LRRs (Kobe and Kajava, 2001), so this SNV is likely not contributing to change in cold growth phenotype and is by chance never seen in the CS isolates in this dataset. The second SNV at position 283800 results in a threonine to methionine change at positon 349, which changes a polar amino acid to a hydrophobic one within a tandem repeat domain. This change is seen in 7 CT isolates (namely A15, A522, A527, A549-2, LM223, LM226, LM78) that are all from serovar 1/2a and 37 INT isolates from serovars 1/2a, 4c and 3a. This SNV may alter protein structure and/or function, but further laboratory tests should be performed to determine the role of this SNV in cold growth. The SNV at position 358406 results in an asparagine to serine substitution at position 424 in lmo0331 (NP_463861.1), another internalin protein. Since these amino acids are both polar and uncharged, this change may not have a significant affect on the protein or on cold growth. This is likely another common SNV that just simply wasn't found in any CS isolates by chance.

Next, the SNV at position 61689 affects lmo0057 (NP_463590.1), which is a hypothetical protein predicted to be from the phage infection protein family and is located adjacent to a heat shock protein lmo0056 (NP_463589.1). 9 CT isolates (namely A10, A15, A522, A527, A549-2, LM119, LM223, LM226, and LM78) from serovar 1/2a and 51 INT isolates from serovars 1/2a and 3a harbour this SNV. This SNV causes a change at position 248 from an alanine, which is hydrophobic, to threonine, which is polar and uncharged. This change could certainly affect this protein to improve growth in cold conditions and should be further studied.

Overall, the CaseControl method identified many SNVs associated with CT isolates versus CS isolates, but were also found in INT growers, that need to be validated *in-vitro* for a causative affect in cold tolerance. I also decided to use another approach, random forests, to discover the most important discriminating SNVs between the CT and

CS groups as a complementary approach since it is better at picking up particular cases that a Fisher's Exact test would miss.

Figure 6.5 shows the top 25 ranked SNV positions as discovered by the random forest based on the mean decrease in accuracy, which indicates the decrease in accuracy of classification of isolates in the phenotypic groups when excluding the specified SNV. Thus, the most important SNVs have the highest mean decrease in accuracy. These important SNVs were also predicted to be significant by the CaseControl method, but some of the less significant SNVs from the first method turned out to be more important using the random forest method.

For example, position 778329 has the highest mean decrease in accuracy, making it the most important SNV in classifying isolates as CT or CS. Looking back at Table 6.1, the p-value association was only 1.89E-03, but this may be a case where we need larger samples to pick up significance with a Fisher Exact test. There are 4 CT isolates (namely A15, A527, A549-2, and LM78) and 26 INT isolates (namely A12, A143, A17, A18, A19, A20, A26, A35, A36, A37, A38, A49, A4, A512, A540, A54, A552, A559, A59, A60, A9, C4, LM109, LM84, LM92, LM93) harbouring this particular SNV. Notably, these isolates are all from serovar 1/2a. This SNV changes position 23 of a hypothetical protein lmo0751 (NP_464278.1) from an alanine, which is hydrophobic, to a glutamic acid, which has a negative charge and is also much larger. Although the structure and function of this protein is unknown, this amino acid change could possibly alter the protein structure and may provide an advantage for cold growth over the CS isolates that do not harbour this change. Further experimental tests are required to validate the role of this SNV in cold growth of serotype 1/2a isolates.

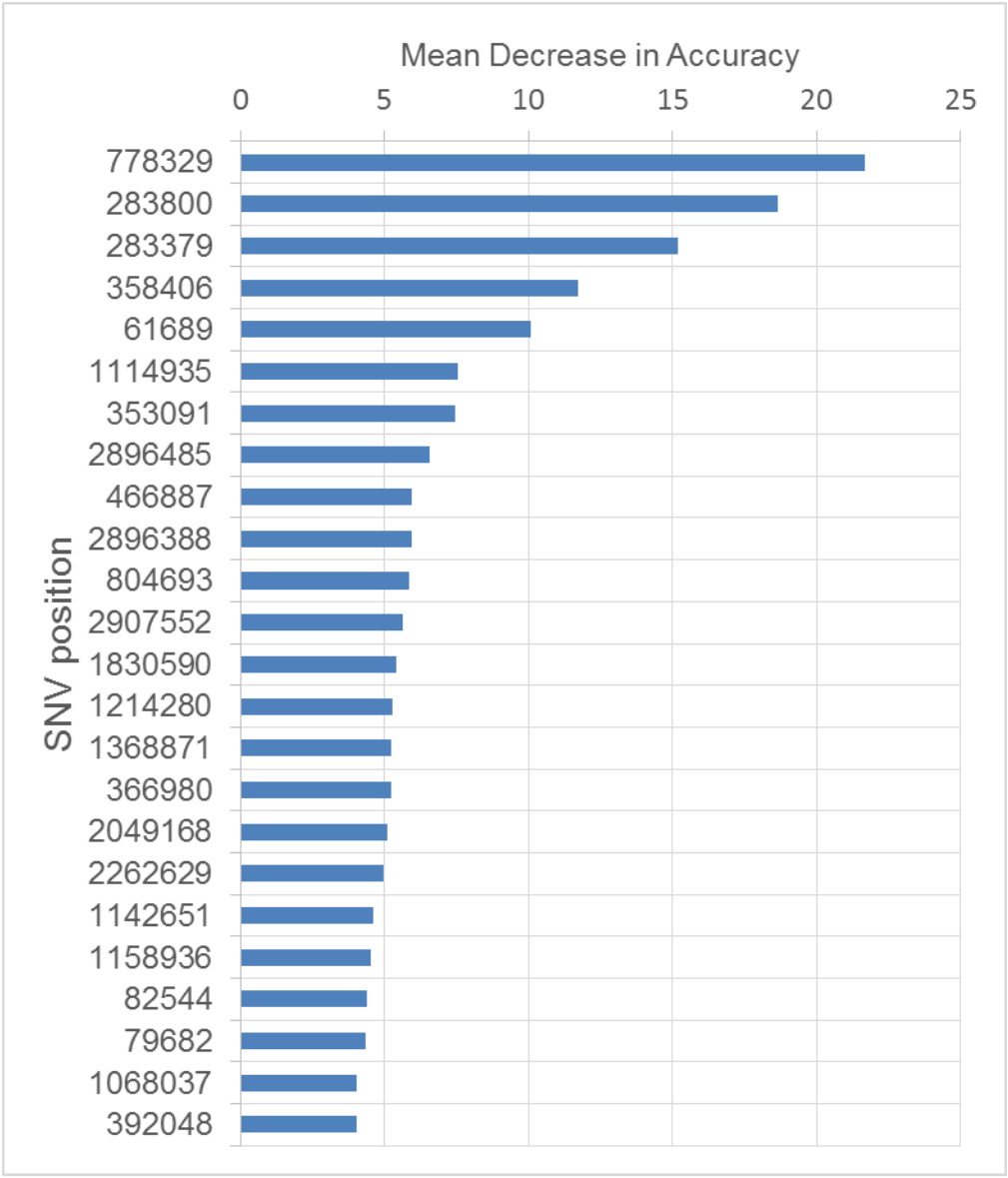

Figure 6.5 Importance, measured by mean decrease in accuracy, of top 25 SNVs used to classify CT and CS isolates as discovered by the random forest method

Overall, these two approaches identified many SNVs associated with CT isolates, that are also present in INT isolates, but never seen in CS isolates. Whether they play a causative role in the adaptation of *L. monocytogenes* to growth in cold temperatures still warrants further investigation through careful selection of SNVs to be tested using *in-vitro* techniques.

6.3.5. SNVs causing sensitivity to cold

In light of the finding that there isn't a clear SNV or set of SNVs in particular genes that may contribute to cold tolerance in *L. monocytogenes*, an alternative hypothesis is that this phenotype is normal behaviour and rather CS isolates have SNVs that impair the regular processes involved in this phenomenon. To further investigate this hypothesis, I identified a set of SNVs unique to the five VCS isolates: A5, LM225, LM231, LM233, and LM96. Table 6.2 summarizes the 98 unique non-synonymous SNVs found in these strains, which includes a number of disruptive variants that are predicted to cause loss of function of genes. Importantly, some of these SNVs also affect major virulence and stress response genes/operons and cell wall/membrane/envelope biogenesis genes, which we would expect to be important for survival in cold temperatures. These variants should be tested further for their impact growth in cold temperatures.

Isolate A5 was collected from a food sample in 1990. In this isolate there are four non-synonymous SNVs predicted to cause a loss of function of a gene either through a premature stop codon or a missing start codon. These include *inlB*, a known VF involved in adhesion and invasion, *rsbV,* another known VF part of the SigB operon involved in general stress response (Chaturongakul and Boor, 2004), lmo1863, a hypothetical protein, and lmo2242, an O6-methylguanine-DNA methyltransferase. Of these, the lack of a functional *rsbV* is a reasonable explanation of the sensitivity of this isolate in responding to cold shock. Two additional small indels were predicted to also cause loss of function of lmo0269, a transporter gene, and lmo1721, a transcriptional regulator. A set of additional missense mutations are also present in predicted AMR genes (e.g. penicillin-binding protein, ultraviolet resistance protein, and multidrug/peptide transporters), and many hypothetical proteins.

Isolate LM231, collected from an asymptomatic human carrier, harboured two non-synonymous SNVs predicted to result in the gain of premature stop codons in *inlA,* one of the most important VFs in *L. monocytogenes*, and lmo1814, a hypothetical protein, gained premature stop codons. The incomplete *inlA* in LM231 is likely the reason this individual was asymptomatic as the bacterium cannot effectively adhere and invade the host cells without this protein (Nightingale et al., 2008). Similar to A5, LM231 also has missense mutations in a number of hypothetical proteins, transporter genes, and other VFs including

lmo2365, which is a transcriptional regulator of the RofA family that has been shown to be involved in the activation of multiple VFs in response to changing environmental conditions in group A streptococci (Beckert et al., 2001). The genes altered in this isolate may also play an important role in the response of *L. monocytogenes* to cold environments.

Another food isolate, LM233, has only two unique non-synonymous SNVs. Lmo0882, a hypothetical protein, gains a premature stop codon and *rsbU* has a missense variant at position 102. *rsbU* is also part of the SigB stress response operon and it is known that N-terminal deletions disrupt its function (Delumeau et al., 2004). Although this strain has an SNV and not a deletion in this region, it is possible that this SNV could also affect the proper function of this stress response regulator and may be causing increased sensitivity in this isolate to cold temperatures.

LM225 and LM96 do not have any unique SNVs that lead to predicted truncation or loss of function in any genes. LM225 only harbours two unique SNVs which impact lmo0479, a predicted secreted protein, and *cobD*, which is involved in cobalamin biosynthesis. It is unclear whether these SNVs are related to cold sensitivity in LM225, and so this suggests there could be regulatory variants or other mechanisms involved in the cold tolerance of *L. monocytogenes*. The sensitivity of LM96 to cold temperatures is likely encoded by 15 unique missense mutations. In addition, a couple of SNVs affect genes involved in DNA replication/recombination/repair, for example *polA* (a DNA polymerase). These could be generally affecting the growth of this isolate even under normal temperatures. There are also other VFs, such as internalins and iron transporters, AMR transporter genes, and hypothetical proteins affected by SNVs in LM96.

In summary, I have identified a set of SNVs unique to VCS isolates in our collection of *L. monocytogenes* that are predicted to impair the function of a set of genes that may be involved in the normal response to cold temperatures.

Table 6.2 Non-synonymous SNVs unique to VCS isolates.

Isolate	SNV position	Reference base	SNV base	Gene	Locus	Accession	Protein change†	Gene product
A5	114898	T	C		lmo0107	NP_463640.1	Asn66Ser	ABC transporter ATP-binding protein
A5	149955	G	C		lmo0152	NP_463685.1	Leu19Val	Peptide ABC transporter substrate-binding protein
A5	205935	G	A	hly	lmo0202	NP_463733.1	Met39Ile	Listeriolysin O precursor
A5	261927	C	T		lmo0241	NP_463772.1	Arg208Cys	Hypothetical protein
A5	341072	C	G		lmo0316	NP_463846.1	Ala43Gly	Hydroxyethylthiazole kinase
A5	396668	C	T		lmo0369	NP_463899.1	Ala154Val	Hypothetical protein
A5	441155	T	C		lmo0420	NP_463949.1	Leu133Ser	Hypothetical protein
A5	457120	C	T	inlB	lmo0434	NP_463963.1	Gln34*	Internalin B
A5	539041	C	T		lmo0503	NP_464031.1	Thr54Ile	PTS fructose transporter subunit IIA
A5	561974	C	T		lmo0526	NP_464054.1	Asp196Asn	Transcriptional regulator
A5	566970	G	A		lmo0529	NP_464057.1	Val403Ile	Glucosaminyltransferase
A5	583340	C	T		lmo0545	NP_464073.1	Gly38Glu	Hypothetical protein
A5	589506	C	T		lmo0551	NP_464079.1	Ser34Phe	Hypothetical protein
A5	618474	C	T		lmo0581	NP_464109.1	Gly124Asp	Hypothetical protein
A5	643450	C	T		lmo0604	NP_464131.1	Gly99Arg	Hypothetical protein
A5	745057	G	C	fliF	lmo0713	NP_464240.1	Ala426Pro	Flagellar MS-ring protein flif
A5	756780	G	A		lmo0727	NP_464254.1	Gly15Ser	Glucosamine--fructose-6-phosphate aminotransferase
A5	929969	T	A	rsbV	lmo0893	NP_464419.1	Tyr26*	Anti-anti-sigma factor
A5	934365	C	T		lmo0898	NP_464424.1	Pro136Leu	Hypothetical protein
A5	1067213	G	A		lmo1036	NP_464561.1	Val273Ile	Hypothetical protein
A5	1170492	C	A		lmo1136	NP_464661.1	Thr164Lys	Internalin
A5	1185540	C	A		lmo1156	NP_464681.1	Leu56Ile	Diol dehydratase-reactivating factor large subunit

Isolate	SNV position	Reference base	SNV base	Gene	Locus	Accession	Protein change†	Gene product
A5	1274654	C	T		lmo1250	NP_464775.1	Gly9Ser	Antibiotic resistance protein
A5	1357225	G	A	truB	lmo1328	NP_464853.1	Asp153Asn	Trna pseudouridine synthase B
A5	1533293	T	A	alaS	lmo1504	NP_465029.1	Met614Leu	Alanyl-trna synthetase
A5	1616031	C	T	dnaE	lmo1574	NP_465099.1	Asp265Asn	DNA polymerase III subunit alpha
A5	1724260	C	A		lmo1670	NP_465195.1	Ala22Asp	Hypothetical protein
A5	1805805	A	T	gltC	lmo1735	NP_465260.1	Lys255Asn	Transcription activator of glutamate synthase operon gltc
A5	1815999	A	T		lmo1746	NP_465271.1	Met375Lys	ABC transporter permease
A5	1841123	C	G	purQ	lmo1769	NP_465294.1	Glu731Gln	Phosphoribosylformylglycinamidine synthase II
A5	1937601	C	T		lmo1863	NP_465388.1	Met1?	Hypothetical protein
A5	2019039	G	A		lmo1943	NP_465467.1	Thr76Ile	Hypothetical protein
A5	2096317	C	T	dapF	lmo2018	NP_465542.1	Asp250Asn	Diaminopimelate epimerase
A5	2112241	G	A		lmo2031	NP_465555.1	Leu45Phe	Hypothetical protein
A5	2121862	G	A	pbpB	lmo2039	NP_465563.1	Pro626Leu	Penicillin-binding protein 2B
A5	2195056	C	A		lmo2115	NP_465639.1	Leu220Ile	ABC transporter permease
A5	2333740	G	A		lmo2242	NP_465766.1	Arg40*	O6-methylguanine-DNA methyltransferase
A5	2538100	G	A		lmo2463	NP_465986.1	Thr161Ile	Multidrug transporter
A5	2566128	C	T	uvrB	lmo2489	NP_466012.1	Asp440Asn	Excinuclease ABC subunit B
A5	2871742	G	T	kat	lmo2785	NP_466307.1	Pro348Gln	Catalase
A5	2907957	G	A		lmo2821	NP_466343.1	Asp269Asn	Internalin
LM225	515481	C	A		lmo0479	NP_464007.1	His31Asn	Secreted protein
LM225	1218937	G	T	cobD	lmo1192	NP_464717.1	Gly255Cys	Cobalamin biosynthesis protein
LM231	93937	C	T		lmo0086	NP_463619.1	Pro1684Ser	Hypothetical protein
LM231	98935	A	G		lmo0090	NP_463623.1	Asp176Gly	ATP synthase F0F1 subunit alpha

Isolate	SNV position	Reference base	SNV base	Gene	Locus	Accession	Protein change†	Gene product
LM231	129653	A	G		lmo0126	NP_463659.1	Ile146Val	Hypothetical protein
LM231	198658	G	A		lmo0195	NP_463726.1	Ser153Asn	ABC transporter permease
LM231	325811	A	G		lmo0300	NP_463831.1	Ile437Thr	Phospho-β-galactosidase
LM231	395605	G	A		lmo0368	NP_463898.1	Glu2Lys	Hypothetical protein
LM231	455509	G	T	inlA	lmo0433	NP_463962.1	Glu326*	Internalin A
LM231	461610	G	A		lmo0435	NP_463964.1	Gly644Ser	Peptidoglycan binding protein
LM231	477207	C	T		lmo0445	NP_463974.1	Ala83Val	Transcriptional regulator
LM231	477431	C	A		lmo0445	NP_463974.1	Gln158Lys	Transcriptional regulator
LM231	530630	A	G		lmo0494	NP_464022.1	Glu39Gly	Hypothetical protein
LM231	543386	C	A	prs	lmo0509	NP_464037.1	Ser62Tyr	Phosphoribosyl pyrophosphate synthetase
LM231	580403	C	A		lmo0541	NP_464069.1	Gly213Cys	ABC transporter substrate-binding protein
LM231	580405	G	T		lmo0541	NP_464069.1	Ala212Glu	ABC transporter substrate-binding protein
LM231	709909	C	T		lmo0674	NP_464201.1	Ala281Thr	Hypothetical protein
LM231	823937	C	G		lmo0796	NP_464323.1	Val78Leu	Hypothetical protein
LM231	1041454	T	A		lmo1012	NP_464537.1	Asn3Lys	N-acyl-L-amino acid amidohydrolase
LM231	1054926	C	T		lmo1026	NP_464551.1	Gly120Glu	Lytr protein
LM231	1533989	T	G	alaS	lmo1504	NP_465029.1	Thr382Pro	Alanyl-trna synthetase
LM231	1648022	A	C		lmo1604	NP_465129.1	Asp115Glu	2-cys peroxiredoxin
LM231	1705543	C	T		lmo1656	NP_465181.1	Ala30Thr	Hypothetical protein
LM231	1805276	A	G	gltC	lmo1735	NP_465260.1	Glu79Gly	Transcription activator of glutamate synthase operon gltc
LM231	1888798	G	T		lmo1814	NP_465339.1	Tyr492*	Hypothetical protein
LM231	1889565	C	A		lmo1814	NP_465339.1	Asp237Tyr	Hypothetical protein
LM231	1963397	A	T		lmo1890	NP_465414.1	Leu70Gln	Hypothetical protein

Isolate	SNV position	Reference base	SNV base	Gene	Locus	Accession	Protein change†	Gene product
LM231	2307987	C	A		lmo2220	NP_465744.1	Gly237Val	3'-5' exoribonuclease
LM231	2357931	G	T	addB	lmo2268	NP_465792.1	Ala623Glu	ATP-dependent deoxyribonuclease subunit B
LM231	2390319	T	G		lmo2312	NP_465836.1	Thr89Pro	Hypothetical protein
LM231	2390345	A	G		lmo2312	NP_465836.1	Val80Ala	Hypothetical protein
LM231	2390350	A	T		lmo2312	NP_465836.1	Asn78Lys	Hypothetical protein
LM231	2392778	G	T		lmo2317	NP_465841.1	Asn273Lys	Hypothetical protein
LM231	2392780	T	C		lmo2317	NP_465841.1	Asn273Asp	Hypothetical protein
LM231	2471700	G	A		lmo2396	NP_465919.1	Asp870Asn	Internalin
LM231	2476612	C	T		lmo2403	NP_465926.1	Glu214Lys	Hypothetical protein
LM231	2481568	A	C		lmo2410	NP_465933.1	Leu3Val	Hypothetical protein
LM231	2551310	C	T		lmo2476	NP_465999.1	Gly329Glu	Aldose 1-epimerase
LM231	2662623	C	T		lmo2582	NP_466105.1	Glu334Lys	Histidine kinase
LM231	2740671	A	T		lmo2667	NP_466189.1	Tyr62Asn	PTS galacticol transporter subunit IIA
LM231	2796195	A	G		lmo2720	NP_466242.1	Glu259Gly	Acetate-coa ligase
LM233	921413	C	T		lmo0882	NP_464408.1	Gln118*	Hypothetical protein
LM233	929042	C	A	rsbU	lmo0892	NP_464418.1	Ala102Glu	Serine phosphatase
LM96	728621	G	T		lmo0694	NP_464221.1	Ala8Ser	Hypothetical protein
LM96	984914	C	A		lmo0950	NP_464475.1	Gly270Val	Hypothetical protein
LM96	1263552	G	A		lmo1236	NP_464761.1	Leu44Phe	Hypothetical protein
LM96	1382303	G	T		lmo1357	NP_464882.1	Arg368Leu	Acetyl-coa carboxylase biotin carboxylase subunit
LM96	1596277	G	T	thrS	lmo1559	NP_465084.1	His291Gln	Threonyl-trna synthetase
LM96	1603620	C	T	polA	lmo1565	NP_465090.1	Met276Ile	DNA polymerase I
LM96	2021357	A	C	resE	lmo1947	NP_465471.1	Val569Gly	Two component sensor histidine kinase
LM96	2194583	C	T		lmo2115	NP_465639.1	Ser62Leu	ABC transporter permease

Isolate	SNV position	Reference base	SNV base	Gene	Locus	Accession	Protein change†	Gene product
LM96	2392744	A	G		lmo2317	NP_465841.1	Phe285Leu	Hypothetical protein
LM96	2500904	G	A		lmo2431	NP_465954.1	Arg83Cys	Ferrichrome ABC transporter substrate-binding protein
LM96	2619190	C	T		lmo2542	NP_466065.1	Gly80Asp	Protophyrinogen oxidase
LM96	2842153	G	A		lmo2760a	YP_008475646.1	Ser40Asn	Hypothetical protein
LM96	2880966	G	C		lmo2794	NP_466316.1	Pro36Ala	NA-binding protein Spo0J
LM96	2907769	G	A		lmo2821	NP_466343.1	Cys206Tyr	Internalin

†Note: * indicates a stop gain, ? indicates a start loss

6.4. Conclusions

In summary, WGS was used to show that no single genetic element or variation was found uniquely in all CT *L. monocytogenes* strains or was even strongly associated with the CT subgroup. Of note, the CT and CS isolates were sporadically distributed in the phylogeny of all isolates and were very closely related in some cases. This would suggest convergent evolution or acquisition of cold-tolerance genes on multiple occasions in the evolution of the cold-tolerance phenotype in *L. monocytogenes*. Since no horizontally-acquired GIs were detected uniquely in the CT group, SNVs may be playing an important role, either in altering genes to promote improved growth in cold conditions, or alternatively, impairing genes involved in the regular response to cold stress. Since the difference in growth curves between CT and INT isolates was not strikingly distinct, this suggests that cold growth could be a normal phenotype for *L. monocytogenes* and the CS isolates simply have impairments in genes that are involved in the regular stress response to cold. So I used two different statistical methods to identify a collection of SNVs observed in isolates that were able to survive and grow at intermediate to high levels after cold shock (INT and CT groups) and never in CS isolates. Furthermore, I also identified a set of SNVs unique to CS isolates that revealed alterations in important stress and virulence related genes that may be contributing to their inability to recover from the stress of being in a cold environment. To assess whether any of these SNVs actually play a causative role in improved tolerance or sensitivity to growth in cold temperatures, further *in-vitro* experiments must be performed, with the end goal in mind of identifying markers that could be used to evaluate the growth potential of *L. monocytogenes* strains found on ready-to-eat foods being stored at cold temperatures.

The cold tolerance phenotype in *L. monocytogenes* is certainly very complex and the work in this chapter reveals that SNVs in the genome alone may not fully explain the phenotypic differences. Previous studies on the transcriptional responses of different strains of *L. monocytogenes* with enhanced or poor growth in cold revealed the importance of the expression of several stress genes, including *sigB*, *cspA*, and *pgpH* (Becker et al., 2000; Liu et al., 2006; Schmid et al., 2009; Arguedas-Villa et al., 2010). However, more recently, a larger study including 62 different strains did not identify any significant

differences in the induction of nine potential cold-adaptation genes (Arguedas-Villa et al., 2014). Whole transcriptome sequencing (or RNAseq) of CT and CS isolates from this collection will be useful to assess differences in the expression of genes between the groups at various time points and may reveal the importance of other mechanisms for survival in cold. In addition, *L. monocytogenes* isolates also exhibit tolerance to other stresses such as desiccation, high salt and high acid concentrations that are often used to control contamination of food products. Patricia Hingston has also performed phenotypic assays of growth under these conditions and I am currently investigating the genetic determinants that may be playing a role in these other phenotypes. Overall, the work in this chapter summarizes the importance of using WGS to better understand the relationship between genotype and phenotype to study characteristics of a pathogen that could be manipulated in the future for better detection of strains that may cause disease.

Chapter 7.

Concluding Remarks

Genome sequencing has seen unprecedented growth since the development of next-generation sequencing technologies that have reduced the cost and time to sequence a genome. Microbial genomes are relatively small and have seen especially high growth in number of sequencing projects in comparison to other organisms. The current standards of data generation can significantly outpace data interpretation and so there is a great need for the development of improved bioinformatics methods for interpretation of microbial genomes in real-time, without sacrificing quality of information.

Upon beginning this project, WGS had only been applied to a handful of infectious disease outbreaks, including a *M. tuberculosis* outbreak in British Columbia that Dr. Brinkman's group studied in collaboration with the BCCDC. I played a role in analyzing the 40 TB genomes to discover that there were no differences in GI content between isolates of this pathogen. However, this application revealed many obstacles in the interpretation of genomes in a timely manner for adding value to existing epidemiological investigations. For one thing, GI prediction was not available for incomplete genomes and had to be done manually through interpretation of genome alignments. In addition to this, there were no methods for the rapid identification of VFs and AMR genes for incomplete genomes either. These types of analyses could reveal important genetic changes in pathogens that could play a role in triggering an outbreak, increasing transmission of the pathogen, or conferring resistance to treatments, and are all essential to study in the context of infectious disease pathogens. Because many of these medically-relevant genes are spread horizontally between diverse organisms via GIs, I aimed to improve characterization of such genes in the framework of GI analysis.

In this dissertation, I first presented my efforts to improve the characterization of microbial genomes, including in the context of virulence and AMR, by expanding the capabilities of IslandViewer, a previously developed web server for the prediction of GIs in microbial genomes. I accomplished this by initially incorporating curated annotations of VFs and AMR genes from quality sources into the genomes available in IslandViewer. To

further improve the coverage of these datasets in closely related strains within a species, I employed a very conservative annotation transfer approach to identify homologs of curated VFs. This step has been integrated into the IslandViewer pipeline so that any new genomes made available after an update will also be analyzed for VFs using this approach. In a similar manner, I used the RGI tool to integrate AMR gene annotations for all genomes in IslandViewer. The command line version of this tool has just recently been released and will also be integrated into the IslandViewer update pipeline for a future release so that new genomes added to the IslandViewer collection will also have AMR gene predictions. In addition to these rich annotations, a final dataset of pathogen-associated genes was also calculated for IslandViewer genomes to quickly highlight those genes that are pathogen-specific and may play an important role in virulence. With the addition of these rich annotation datasets, I also played a role in improving the visual interface of IslandViewer in collaboration with Matthew Laird to provide more interactive and dynamic evaluation of a given genome. Custom genome analysis through IslandViewer's predictive pipeline is of utmost importance to support the use of this tool for the study of newly sequenced genomes. In this regard, I played a collaborative role with Julie Shay to integrate the analysis of incomplete genomes in IslandViewer. In the near future, the developed protocols for identification of VF and AMR gene homologs will also be integrated into the custom genomes analysis pipeline. Overall, this portion of my thesis project has helped improve characterization of VFs, AMR genes and pathogen-associated genes for microbial genomes in IslandViewer for immediate interpretation both in the context of GIs and in general across the rest of the genome.

I utilized these novel developments in IslandViewer to conduct two analyses of trends of GIs. First, I took advantage of the recently integrated AMR gene predictions and coupled them with the GI predictions for all genomes in IslandViewer to perform the first comprehensive analysis of the association of AMR genes with mobile elements. Previous studies have shown many AMR genes are spread horizontally via GIs and plasmids, but this association had never been tested over a collection of diverse genera. The analysis I performed revealed that AMR genes are in fact not disproportionately associated with GIs or plasmids, and instead are found at higher levels on the rest of the chromosome, which further supports the ancient origin theory of AMR. What's more, the high levels of AMR genes on the chromosome were not explained simply by the association of mutations that

confer resistance. Upon separation of the AMR genes based on resistance classes, trends were observed in classes that were disproportionately associated with mobile elements, such as aminoglycoside and β-lactam resistance, that would be considered high risk for spreading, and those that were disproportionately associated with the rest of the chromosome, including fluoroquinolone and peptide antibiotics, that can be treated as lower risk of spreading. Certain mechanisms of action of antibiotic resistances were also disproportionately associated with mobile elements or not. Efflux pumps, for example, were not associated with mobile elements and instead are found at higher levels on the rest of the chromosome, which also supports the notion that such resistance machinery is anciently developed. In summary, the characterization of AMR genes over a diverse collection of genomes coupled with GI predictions using IslandViewer provides a better understanding of the evolution of resistance in microbial species as well as the trends in mobility of the various classes of AMR that can be used to inform risk assessment.

Secondly, I performed the first analysis of the dynamics of GIs across multiple outbreaks of *L. monocytogenes* from across Canada using IslandViewer to show that each geographically and temporally separate outbreak had unique GIs. This is likely linked to the separated exposure of outbreak isolates from different locations to mobile elements in the environment, whether that is from phage or other sources. By no means is this a comprehensive nor representative evaluation of the behaviour of GIs in any type of infectious disease outbreak, but rather it provides a deeper understanding of GIs in *L. monocytogenes* over time and geographic separation and shows the importance of evaluating such data in genomic studies.

Finally, the final portion of my dissertation presents an additional genomic analysis of *L. monocytogenes*, in this case, to determine the genetic determinants specifically responsible for the cold-tolerance phenotype. The ability of *L. monocytogenes* to survive in cold temperatures is a major problem for the food industry to control the levels of growth of this pathogen in contaminated ready-to-eat cold foods. This portion of my thesis focused on detecting SNVs and/or large scale GI acquisitions that were shared by all cold-tolerant isolates that could explain this phenotype, however, no such elements were detected. A collection of SNVs found in subsets of cold-tolerant isolates and intermediate growers, but never in cold-sensitive isolates, was compiled. Notably, the CT isolates were

phylogenetically sporadic and were closely related to CS isolates in some cases. Instead, I suggest that this phenotype could be normal for *L. monocytogenes*, especially since the gradient of growth curve measurements was not strikingly distinct between the groups. In this case, the severely cold-sensitive isolates instead have impairments in genes that are involved in the regular response to this stress. Thus, I was able to identify SNVs in cold-sensitive isolates that impaired function of a number of genes that warrant further investigation for their role in the cold tolerance phenotype of *L. monocytogenes*.

In conclusion, this dissertation portrays improved characterization of GIs and mobile elements in microbial genomes and utilized this information to conduct two novel analyses of trends in the mobilization of AMR genes and the spread of GIs in *L. monocytogenes* outbreaks. The tools developed here will be key assets for future research of microbial genomes as the field moves towards analyzing hundreds to thousands of genomes and will act as a base for expanded development of IslandViewer and other related tools. This work builds upon our knowledge of mobile elements in microbial genomes and can enrich future studies, particularly in the field of infectious diseases.

References

Aarestrup, F.M., Brown, E.W., Detter, C., Gerner-Smidt, P., Gilmour, M.W., Harmsen, D., Hendriksen, R.S., Hewson, R., Heymann, D.L., and Johansson, K. (2012). Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. Emerging Infectious Diseases (Print Edition) *18,* e1.

Abedon, S.T., and Lejeune, J.T. (2007). Why bacteriophage encode exotoxins and other virulence factors. Evol. Bioinform Online *1,* 97-110.

Abram, F., Starr, E., Karatzas, K.A., Matlawska-Wasowska, K., Boyd, A., Wiedmann, M., Boor, K.J., Connally, D., and O'Byrne, C.P. (2008). Identification of components of the sigma B regulon in Listeria monocytogenes that contribute to acid and salt tolerance. Appl. Environ. Microbiol. *74,* 6848-6858.

Alton, N.K., and Vapnek, D. (1979). Nucleotide sequence analysis of the chloramphenicol resistance transposon Tn9. Nature *282,* 864.

Alvarez-Martinez, C.E., and Christie, P.J. (2009). Biological diversity of prokaryotic type IV secretion systems. Microbiol. Mol. Biol. Rev. *73,* 775-808.

Aminov, R.I. (2009). The role of antibiotics and antibiotic resistance in nature. Environ. Microbiol. *11,* 2970-2988.

Angiuoli, S.V., and Salzberg, S.L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics *27,* 334-342.

Arca, P., Rico, M., Brana, A.F., Villar, C.J., Hardisson, C., and Suarez, J.E. (1988). Formation of an adduct between fosfomycin and glutathione: a new mechanism of antibiotic resistance in bacteria. Antimicrob. Agents Chemother. *32,* 1552-1556.

Arguedas-Villa, C., Stephan, R., and Tasara, T. (2010). Evaluation of cold growth and related gene transcription responses associated with Listeria monocytogenes strains of different origins. Food Microbiol. *27,* 653-660.

Arguedas-Villa, C., Kovacevic, J., Allen, K.J., Stephan, R., and Tasara, T. (2014). Cold growth behaviour and genetic comparison of Canadian and Swiss Listeria monocytogenes strains associated with the food supply chain and human listeriosis cases. Food Microbiol. *40,* 81-87.

Arias, C.A., Panesso, D., McGrath, D.M., Qin, X., Mojica, M.F., Miller, C., Diaz, L., Tran, T.T., Rincon, S., and Barbu, E.M. (2011). Genetic basis for in vivo daptomycin resistance in enterococci. N. Engl. J. Med. *365,* 892-900.

Armstrong, R.W., Kurucsev, T., and Strauss, U.P. (1970). Interaction between acridine dyes and deoxyribonucleic acid. J. Am. Chem. Soc. *92,* 3174-3181.

Arthur, M., Reynolds, P., Depardieu, F., Evers, S., Dutka-Malen, S., Quintiliani, R., and Courvalin, P. (1996). Mechanisms of glycopeptide resistance in enterococci. J. Infect. *32,* 11-16.

Arthur, M., and Courvalin, P. (1993). Genetics and mechanisms of glycopeptide resistance in enterococci. Antimicrob. Agents Chemother. *37,* 1563-1571.

Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ *3,* e1029.

Aubry-Damon, H., Soussy, C.J., and Courvalin, P. (1998). Characterization of mutations in the rpoB gene that confer rifampin resistance in Staphylococcus aureus. Antimicrob. Agents Chemother. *42,* 2590-2594.

Baar, C., Eppinger, M., Raddatz, G., Simon, J., Lanz, C., Klimmek, O., Nandakumar, R., Gross, R., Rosinus, A., Keller, H.*, et al.* (2003). Complete genome sequence and analysis of Wolinella succinogenes. Proc. Natl. Acad. Sci. U. S. A. *100,* 11690-11695.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., and Prjibelski, A.D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology *19,* 455-477.

Banks, D.J., Beres, S.B., and Musser, J.M. (2002). The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. Trends Microbiol. *10,* 515-521.

Barbosa, W.B., Cabedo, L., Wederquist, H.J., Sofos, J.N., and Schmidt, G.R. (1994). Growth variation among species and strains of Listeria in culture broth. Journal of Food Protection® *57,* 765-769.

Barlow, M., and Hall, B.G. (2002). Phylogenetic analysis shows that the OXA b-lactamase genes have been on plasmids for millions of years. J. Mol. Evol. *55,* 314-321.

Bashir, A., Klammer, A.A., Robins, W.P., Chin, C., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., and Peluso, P. (2012). A hybrid approach for the automated finishing of bacterial genomes. Nat. Biotechnol. *30,* 701-707.

Becavin, C., Bouchier, C., Lechat, P., Archambaud, C., Creno, S., Gouin, E., Wu, Z., Kuhbacher, A., Brisse, S., Pucciarelli, M.G.*, et al.* (2014). Comparison of widely used Listeria monocytogenes strains EGD, 10403S, and EGD-e highlights genomic variations underlying differences in pathogenicity. MBio *5,* e00969-14.

Beckert, S., Kreikemeyer, B., and Podbielski, A. (2001). Group A streptococcal rofA gene is involved in the control of several virulence genes and eukaryotic cell attachment and internalization. Infect. Immun. *69,* 534-537.

Bennett, P. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. Br. J. Pharmacol. *153,* S347-S357.

Bergholz, T.M., den Bakker, H.C., Fortes, E.D., Boor, K.J., and Wiedmann, M. (2010). Salt stress phenotypes in Listeria monocytogenes vary by genetic lineage and temperature. Foodborne Pathogens and Disease *7,* 1537-1549.

Bergholz, T.M., den Bakker, H.C., Katz, L.S., Silk, B.J., Jackson, K.A., Kucerova, Z., Joseph, L.A., Turnsek, M., Gladney, L.M., Halpin, J.L.*, et al.* (2015). Determination of Evolutionary Relationships of Outbreak-Associated Listeria monocytogenes Strains of Serotypes 1/2a and 1/2b by Whole-Genome Sequencing. Appl. Environ. Microbiol. *82,* 928-938.

Bhullar, K., Waglechner, N., Pawlowski, A., Koteva, K., Banks, E.D., Johnston, M.D., Barton, H.A., and Wright, G.D. (2012). Antibiotic resistance is prevalent in an isolated cave microbiome. PLoS One *7,* e34953.

Bierne, H., Sabet, C., Personnic, N., and Cossart, P. (2007). Internalins: a complex family of leucine-rich repeat-containing proteins in Listeria monocytogenes. Microb. Infect. *9,* 1156-1166.

Blahna, M.T., Zalewski, C.A., Reuer, J., Kahlmeter, G., Foxman, B., and Marrs, C.F. (2006). The role of horizontal gene transfer in the spread of trimethoprim-sulfamethoxazole resistance among uropathogenic Escherichia coli in Europe and Canada. J. Antimicrob. Chemother. *57,* 666-672.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30,* 2114-2120.

Bonnet, R. (2004). Growing group of extended-spectrum beta-lactamases: the CTX-M enzymes. Antimicrob. Agents Chemother. *48,* 1-14.

Boucher, Y., Labbate, M., Koenig, J.E., and Stokes, H. (2007). Integrons: mobilizable platforms that promote genetic diversity in bacteria. Trends Microbiol. *15,* 301-309.

Boyd, D.A., Peters, G.A., Ng, L., and Mulvey, M.R. (2000). Partial characterization of a genomic island associated with the multidrug resistance region of Salmonella enterica Typhymurium DT104. FEMS Microbiol. Lett. *189,* 285-291.

Boyd, E.F., Davis, B.M., and Hochhut, B. (2001). Bacteriophage–bacteriophage interactions in the evolution of pathogenic bacteria. Trends Microbiol. *9,* 137-144.

Boyd, A.P., Ross, P.J., Conroy, H., Mahon, N., Lavelle, E.C., and Mills, K.H. (2005). Bordetella pertussis adenylate cyclase toxin modulates innate and adaptive immune responses: distinct roles for acylation and enzymatic activity in immunomodulation and cell death. J. Immunol. *175,* 730-738.

Braun, L., Ohayon, H., and Cossart, P. (1998). The InlB protein of Listeria monocytogenes is sufficient to promote entry into mammalian cells. Mol. Microbiol. *27,* 1077-1087.

Breiman, L. (2001). Random forests. Mach. Learning *45,* 5-32.

Bulinski, A., Butkovsky, O., Shashkin, A., and Yaskov, P. (2011). Statistical methods of SNP data analysis with applications. arXiv Preprint arXiv:1106.4989

Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P., and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. Genet. Epidemiol. *28,* 171-182.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10,* 1.

Campbell, A., Schneider, S., and Song, B. (1992). Lambdoid phages as elements of bacterial genomes (integrase/phage21/Escherichia coli K-12/icd gene). Genetica *86,* 259-267.

Canchaya, C., Fournous, G., and Brüssow, H. (2004). The impact of prophages on bacterial chromosomes. Mol. Microbiol. *53,* 9-18.

Carattoli, A. (2003). Plasmid-mediated antimicrobial resistance in Salmonella enterica. Curr. Issues Mol. Biol. *5,* 113-122.

Cartwright, E.J., Jackson, K.A., Johnson, S.D., Graves, L.M., Silk, B.J., and Mahon, B.E. (2013). Listeriosis outbreaks and associated food vehicles, United States, 1998-2008. Emerg. Infect. Dis. *19,* 1-9; quiz 184.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G., and Parkhill, J. (2005). ACT: the Artemis Comparison Tool. Bioinformatics *21,* 3422-3423.

Casadevall, A., and Pirofski, L. (2001). Host-pathogen interactions: the attributes of virulence. J. Infect. Dis. *184,* 337-344.

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

Casjens, S., Palmer, N., Van Vugt, R., Mun Huang, W., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., and Dodson, R.J. (2000). A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete Borrelia burgdorferi. Mol. Microbiol. *35,* 490-516.

Chakraborty, T., Hain, T., and Domann, E. (2000). Genome organization and the evolution of the virulence gene locus in Listeria species. International Journal of Medical Microbiology *290,* 167-174.

Chandrasekaran, S., and Lalithakumari, D. (1998). Plasmid-mediated rifampicin resistance in Pseudomonas fluorescens. J. Med. Microbiol. *47,* 197-200.

Charpentier, E., Gerbaud, G., and Courvalin, P. (1993). Characterization of a new class of tetracycline-resistance gene tet (S) in Listeria monocytogenes BM4210. Gene *131,* 27-34.

Charpentier, E., and Courvalin, P. (1997). Emergence of the trimethoprim resistance gene dfrD in Listeria monocytogenes BM4293. Antimicrob. Agents Chemother. *41,* 1134-1136.

Chaturongakul, S., and Boor, K.J. (2004). RsbT and RsbV contribute to sigmaB-dependent survival under environmental, energy, and intracellular stress conditions in Listeria monocytogenes. Appl. Environ. Microbiol. *70,* 5349-5356.

Che, D., Hasan, M.S., Wang, H., Fazekas, J., Huang, J., and Liu, Q. (2011). EGID: an ensemble algorithm for improved genomic island detection in genomic sequences. Bioinformation *7,* 311-314.

Cheetham, B.F., and Katz, M.E. (1995). A role for bacteriophages in the evolution and transfer of bacterial virulence determinants. Mol. Microbiol. *18,* 201-208.

Chen, L., Xiong, Z., Sun, L., Yang, J., and Jin, Q. (2012). VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res. *40,* D641-D645.

Chen, P.E., and Shapiro, B.J. (2015). The advent of genome-wide association studies for bacteria. Curr. Opin. Microbiol. *25,* 17-24.

Chen, Y., Ross, W.H., Scott, V.N., and Gombas, D.E. (2003). Listeria monocytogenes: low levels equal low risk. Journal of Food Protection® *66,* 570-577.

Chen, L., Zheng, D., Liu, B., Yang, J., and Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. Nucleic Acids Res. *44,* D694-7.

Christie, P.J. (2001). Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. Mol. Microbiol. *40,* 294-305.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311,* 1283-1287.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6,* 80-92.

Clatworthy, A.E., Pierson, E., and Hung, D.T. (2007). Targeting virulence: a new paradigm for antimicrobial therapy. Nature Chemical Biology *3,* 541-548.

Cloeckaert, A., Baucheron, S., and Chaslus-Dancla, E. (2001). Nonenzymatic chloramphenicol resistance mediated by IncC plasmid R55 is encoded by a floR gene variant. Antimicrob. Agents Chemother. *45,* 2381-2382.

Cohen, S.P., Hachler, H., and Levy, S.B. (1993). Genetic and functional analysis of the multiple antibiotic resistance (mar) locus in Escherichia coli. J. Bacteriol. *175,* 1484-1492.

Cohen, S.P., McMurry, L.M., Hooper, D.C., Wolfson, J.S., and Levy, S.B. (1989). Cross-resistance to fluoroquinolones in multiple-antibiotic-resistant (Mar) Escherichia coli selected by tetracycline or chloramphenicol: decreased drug accumulation associated with membrane changes in addition to OmpF reduction. Antimicrob. Agents Chemother. *33,* 1318-1325.

Cossart, P. (2011). Illuminating the landscape of host-pathogen interactions with the bacterium Listeria monocytogenes. Proc. Natl. Acad. Sci. U. S. A. *108,* 19484-19491.

Cossart, P., Vicente, M.F., Mengaud, J., Baquero, F., Perez-Diaz, J.C., and Berche, P. (1989). Listeriolysin O is essential for virulence of Listeria monocytogenes: direct evidence obtained by gene complementation. Infect. Immun. *57,* 3629-3636.

Courvalin, P., and Calier, C. (1981). Resistance towards aminoglycoside-aminocyclitol antibiotics in bacteria. J. Antimicrob. Chemother. *8,* 57-69.

Courvalin, P. (2006). Vancomycin resistance in gram-positive cocci. Clin. Infect. Dis. *42 Suppl 1,* S25-34.

Courvalin, P. (1994). Transfer of antibiotic resistance genes between gram-positive and gram-negative bacteria. Antimicrob. Agents Chemother. *38,* 1447-1451.

Croucher, N.J., Mostowy, R., Wymant, C., Turner, P., Bentley, S.D., and Fraser, C. (2016). Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. PLoS Biol *14,* e1002394.

D'Costa, V.M., King, C.E., Kalan, L., Morar, M., Sung, W.W., Schwarz, C., Froese, D., Zazula, G., Calmels, F., and Debruyne, R. (2011). Antibiotic resistance is ancient. Nature *477,* 457-461.

Dabbs, E.R., Yazawa, K., Tanaka, Y., Mikami, Y., Miyaji, M., Andersen, S.J., Morisaki, N., Iwasaki, S., Shida, O., and Takagi, H. (1995a). Rifampicin inactivation by Bacillus species. J. Antibiot. *48,* 815-819.

Dabbs, E.R., Yazawa, K., Mikami, Y., Miyaji, M., Morisaki, N., Iwasaki, S., and Furihata, K. (1995b). Ribosylation by mycobacterial strains as a new mechanism of rifampin inactivation. Antimicrob. Agents Chemother. *39,* 1007-1009.

Dalgaard, P., and Koutsoumanis, K. (2001). Comparison of maximum specific growth rates and lag times estimated from absorbance and viable count data by different mathematical models. J. Microbiol. Methods *43,* 183-196.

Darling, A.C.E., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. *14,* 1394.

Datta, N., and Hedges, R. (1972). Trimethoprim resistance conferred by W plasmids in Enterobacteriaceae. Microbiology *72,* 349-355.

Davies, J., and Davies, D. (2010). Origins and evolution of antibiotic resistance. Microbiol. Mol. Biol. Rev. *74,* 417-433.

D'Costa, V.M., McGrann, K.M., Hughes, D.W., and Wright, G.D. (2006). Sampling the antibiotic resistome. Science *311,* 374-377.

de Brito, D.M., Maracaja-Coutinho, V., de Farias, S.T., Batista, L.V., and do Rêgo, T.G. (2016). A Novel Method to Predict Genomic Islands Based on Mean Shift Clustering Algorithm. PloS One *11,* e0146352.

De Lobel, L., Geurts, P., Baele, G., Castro-Giner, F., Kogevinas, M., and Van Steen, K. (2010). A screening methodology based on Random Forests to improve the detection of gene–gene interactions. European Journal of Human Genetics *18,* 1127-1132.

Delumeau, O., Dutta, S., Brigulla, M., Kuhnke, G., Hardwick, S.W., Volker, U., Yudkin, M.D., and Lewis, R.J. (2004). Functional and structural characterization of RsbU, a stress signaling protein phosphatase 2C. J. Biol. Chem. *279,* 40927-40937.

Dhillon, B.K., Chiu, T.A., Laird, M.R., Langille, M.G., and Brinkman, F.S. (2013). IslandViewer update: Improved genomic island discovery and visualization. Nucleic Acids Res. *41,* W129-32.

Dhillon, B.K., Laird, M.R., Shay, J.A., Winsor, G.L., Lo, R., Nizam, F., Pereira, S.K., Waglechner, N., McArthur, A.G., Langille, M.G., and Brinkman, F.S. (2015). IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic Acids Res. *43,* W104-8.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. Nature Reviews Microbiology *2,* 414-424.

Doi, Y., and Arakawa, Y. (2007). 16S ribosomal RNA methylation: emerging resistance mechanism against aminoglycosides. Clin. Infect. Dis. *45,* 88-94.

Durack, J., Ross, T., and Bowman, J.P. (2013). Characterisation of the transcriptomes of genetically diverse Listeria monocytogenes exposed to hyperosmotic and low temperature conditions reveal global stress-adaptation mechanisms. PloS One *8,* e73603.

Dutilh, B.E., Backus, L., Edwards, R.A., Wels, M., Bayjanov, J.R., and van Hijum, S.A. (2013). Explaining microbial phenotypes on a genomic scale: GWAS for microbes. Brief Funct. Genomics *12,* 366-380.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput Biol *7,* e1002195.

Elhai, J., Liu, H., and Taton, A. (2012). Detection of horizontal transfer of individual genes by anomalous oligomer frequencies. BMC Genomics *13,* 1.

Elwell, L.P., and Shipley, P.L. (1980). Plasmid-mediated factors associated with virulence of bacteria to animals. Annual Reviews in Microbiology *34,* 465-496.

Emond-Rheault, J., Vincent, A.T., Trudel, M.V., Brochu, F., Boyle, B., Tanaka, K.H., Attéré, S.A., Jubinville, É, Loch, T.P., and Winters, A.D. (2015). Variants of a genomic island in Aeromonas salmonicida subsp. salmonicida link isolates with their geographical origins. Vet. Microbiol. *175,* 68-76.

Epand, R.M., and Vogel, H.J. (1999). Diversity of antimicrobial peptides and their mechanisms of action. Biochimica Et Biophysica Acta (BBA)-Biomembranes *1462,* 11-28.

Falkow, S. (2004). Molecular Koch's postulates applied to bacterial pathogenicity—a personal recollection 15 years later. Nature Reviews Microbiology *2,* 67-72.

Falkow, S. (1988). Molecular Koch's postulates applied to microbial pathogenicity. Review of Infectious Diseases *10,* S274-S276.

Falush, D. (2009). Toward the use of genomics to study microevolutionary change in bacteria. PLoS Genetics *5,* e1000627.

Falush, D., and Bowden, R. (2006). Genome-wide association mapping in bacteria? Trends Microbiol. *14,* 353-355.

Farhat, M.R., Shapiro, B.J., Kieser, K.J., Sultana, R., Jacobson, K.R., Victor, T.C., Warren, R.M., Streicher, E.M., Calver, A., and Sloutsky, A. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat. Genet. *45,* 1183-1189.

Feil, E.J. (2004). Small change: keeping pace with microevolution. Nature Reviews Microbiology *2,* 483-495.

Fermer, C., Kristiansen, B.E., Sköld, O., and Swedberg, G. (1995). Sulfonamide resistance in Neisseria meningitidis as defined by site-directed mutagenesis could have its origin in other species. J. Bacteriol. *177,* 4669-4675.

Fernández, X.M., and Birney, E. (2010). Ensembl Genome Browser. In Vogel and Motulsky's Human Genetics, Springer) pp. 923-939.

Fines, M., and Leclercq, R. (2000). Activity of linezolid against Gram-positive cocci possessing genes conferring resistance to protein synthesis inhibitors. J. Antimicrob. Chemother. *45,* 797-802.

Finley, R.L., Collignon, P., Larsson, D.G., McEwen, S.A., Li, X.Z., Gaze, W.H., Reid-Smith, R., Timinouni, M., Graham, D.W., and Topp, E. (2013). The scourge of antibiotic resistance: the important role of the environment. Clin. Infect. Dis. *57,* 704-710.

Fleming, D.W., Cochi, S.L., MacDonald, K.L., Brondum, J., Hayes, P.S., Plikaytis, B.D., Holmes, M.B., Audurier, A., Broome, C.V., and Reingold, A.L. (1985). Pasteurized milk as a vehicle of infection in an outbreak of listeriosis. N. Engl. J. Med. *312,* 404-407.

Fonseca, N.A., Rung, J., Brazma, A., and Marioni, J.C. (2012). Tools for mapping high-throughput sequencing data. Bioinformatics *28,* 3169-3177.

Fortier, L., and Sekulovic, O. (2013). Importance of prophages to evolution and virulence of bacterial pathogens. Virulence *4,* 354-365.

Foster, T.J., Davis, M.A., Roberts, D.E., Takeshita, K., and Kleckner, N. (1981). Genetic organization of transposon Tn10. Cell *23,* 201-213.

Fournier, P.E., Drancourt, M., and Raoult, D. (2007). Bacterial genome sequencing and its use in infectious diseases. The Lancet Infectious Diseases *7,* 711-723.

Fournier, P., Vallenet, D., Barbe, V., Audic, S., Ogata, H., Poirel, L., Richet, H., Robert, C., Mangenot, S., and Abergel, C. (2006). Comparative genomics of multidrug resistance in Acinetobacter baumannii. PLoS Genet *2,* e7.

Gaillard, J., Berche, P., Frehel, C., Gouln, E., and Cossart, P. (1991). Entry of L. monocytogenes into cells is mediated by internalin, a repeat protein reminiscent of surface antigens from gram-positive cocci. Cell *65,* 1127-1141.

Garcia-Lobo, J.M., and Ortiz, J.M. (1982). Tn292l, a transposon encoding fosfomycin resistance. J. Bacteriol. *151,* 477-479.

Garcia-Vallve, S., Guzman, E., Montero, M.A., and Romeu, A. (2003). HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. *31,* 187-189.

Garcia-Vallve, S., Romeu, A., and Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. *10,* 1719-1725.

Gardy, J.L., Johnston, J.C., Sui, S.J.H., Cook, V.J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., and Holt, R. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N. Engl. J. Med. *364,* 730-739.

Garg, A., and Gupta, D. (2008). VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics *9,* 1.

Gibreel, A., and Sköld, O. (1999). Sulfonamide resistance in clinical isolates of Campylobacter jejuni: mutational changes in the chromosomal dihydropteroate synthase. Antimicrob. Agents Chemother. *43,* 2156-2160.

Gill, E.E., and Brinkman, F.S. (2011). The proportional lack of archaeal pathogens: Do viruses/phages hold the key? Bioessays *33,* 248-254.

Gillings, M.R., Krishnan, S., Worden, P.J., and Hardwick, S.A. (2008). Recovery of diverse genes for class 1 integron-integrases from environmental DNA samples. FEMS Microbiol. Lett. *287,* 56-62.

Gilmour, M.W., Graham, M., Van Domselaar, G., Tyler, S., Kent, H., Trout-Yakel, K.M., Larios, O., Allen, V., Lee, B., and Nadon, C. (2010). High-throughput genome sequencing of two Listeria monocytogenes clinical isolates during a large foodborne outbreak. BMC Genomics *11,* 120-2164-11-120.

Gouin, E., Adib-Conquy, M., Balestrino, D., Nahori, M.A., Villiers, V., Colland, F., Dramsi, S., Dussurget, O., and Cossart, P. (2010). The Listeria monocytogenes InlC protein interferes with innate immune responses by targeting the I{kappa}B kinase subunit IKK{alpha}. Proc. Natl. Acad. Sci. U. S. A. *107,* 17333-17338.

Griffith, F. (1928). The significance of pneumococcal types. J. Hyg. *27,* 113-159.

Groisman, E.A., and Ochman, H. (1996). Pathogenicity islands: bacterial evolution in quantum leaps. Cell *87,* 791-794.

Gunn, J.S., Lim, K.B., Krueger, J., Kim, K., Guo, L., Hackett, M., and Miller, S.I. (1998). PmrA–PmrB-regulated genes necessary for 4-aminoarabinose lipid A modification and polymyxin resistance. Mol. Microbiol. *27,* 1171-1182.

Gupta, N., Limbago, B.M., Patel, J.B., and Kallen, A.J. (2011). Carbapenem-resistant Enterobacteriaceae: epidemiology and prevention. Clin. Infect. Dis. *53,* 60-67.

Gupta, S.K., Padmanabhan, B.R., Diene, S.M., Lopez-Rojas, R., Kempf, M., Landraud, L., and Rolain, J.M. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob. Agents Chemother. *58,* 212-220.

Guttman, B., Raya, R., and Kutter, E. (2004). 3 Basic Phage Biology. Bacteriophages: Biology and Applications 29.

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

Hacker, J., and Kaper, J.B. (2000). Pathogenicity islands and the evolution of microbes. Annual Reviews in Microbiology *54,* 641-679.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R., and Goebel, W. (1990). Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extra intestinal Escherichia coli isolates. Microb. Pathog. *8,* 213-225.

Hacker, J., and Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. EMBO Rep. *2,* 376-381.

Haft, D.H., Selengut, J.D., Brinkac, L.M., Zafar, N., and White, O. (2005). Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. Bioinformatics *21,* 293-306.

Hagens, S., and Loessner, M.J. (2015). Phages of Listeria offer novel tools for diagnostics and biocontrol. Gram-Positive Phages: From Isolation to Application 94.

Hall, R.M., and Collis, C.M. (1998). Antibiotic resistance in gram-negative bacteria: the role of gene cassettes and integrons. Drug Resistance Updates *1,* 109-119.

Hamano, Y., Matsuura, N., Kitamura, M., and Takagi, H. (2006). A novel enzyme conferring streptothricin resistance alters the toxicity of streptothricin D from broad-spectrum to bacteria-specific. J. Biol. Chem. *281,* 16842-16848.

Hanks, M.C., Newman, B., Oliver, I.R., and Masters, M. (1988). Packaging of transducing DNA by bacteriophage P1. Molecular and General Genetics MGG *214,* 523-532.

Harris, S.R., Feil, E.J., Holden, M.T.G., Quail, M.A., Nickerson, E.K., Chantratita, N., Gardete, S., Tavares, A., Day, N., and Lindsay, J.A. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. Science *327,* 469.

Hartman, B.J., and Tomasz, A. (1984). Low-affinity penicillin-binding protein associated with beta-lactam resistance in Staphylococcus aureus. J. Bacteriol. *158,* 513-516.

Hasan, M.S., Liu, Q., Wang, H., Fazekas, J., Chen, B., and Che, D. (2012). GIST: Genomic island suite of tools for predicting genomic islands in genomic sequences. Bioinformation *8,* 203-205.

Hatem, A., Bozdag, D., Toland, A.E., and Catalyurek, U.V. (2013). Benchmarking short sequence mapping tools. BMC Bioinformatics *14,* 184-2105-14-184.

Haubold, B., and Wiehe, T. (2006). How repetitive are genomes? BMC Bioinformatics *7,* 541.

He, Z., Zhang, H., Gao, S., Lercher, M.J., Chen, W.H., and Hu, S. (2016). Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res.

Health Canada. (2011). Policy on Listeria monocytogenes in Ready-to-Eat foods.

Heep, M., Beck, D., Bayerdorffer, E., and Lehn, N. (1999). Rifampin and rifabutin resistance mechanism in Helicobacter pylori. Antimicrob. Agents Chemother. *43,* 1497-1499.

Hegyi, H., and Gerstein, M. (2001). Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. Genome Res. *11,* 1632-1640.

Helweg-Larsen, J., Benfield, T.L., Eugen-Olsen, J., Lundgren, J.D., and Lundgren, B. (1999). Effects of mutations in Pneumocystis carinii dihydropteroate synthase gene on outcome of AIDS-associated P carinii pneumonia. The Lancet *354,* 1347-1351.

Hensel, M., Shea, J.E., Gleeson, C., Jones, M.D., Dalton, E., and Holden, D.W. (1995). Simultaneous identification of bacterial virulence genes by negative selection. Science *269,* 400-403.

Ho Sui, S.J., Fedynak, A., Hsiao, W.W., Langille, M.G., and Brinkman, F.S. (2009). The association of virulence factors with genomic islands. PLoS One *4,* e8094.

Hochhut, B., Jahreis, K., Lengeler, J.W., and Schmid, K. (1997). CTnscr94, a conjugative transposon found in enterobacteria. J. Bacteriol. *179,* 2097-2102.

Holden, M., Crossman, L., Cerdeño-Tárraga, A., and Parkhill, J. (2004). Pathogenomics of non-pathogens. Nature Reviews Microbiology *2,* 91-91.

Hooper, D.C. (2001). Emerging mechanisms of fluoroquinolone resistance. Emerg. Infect. Dis. *7,* 337-341.

Horinouchi, S., and Weisblum, B. (1982). Nucleotide sequence and functional map of pC194, a plasmid that specifies inducible chloramphenicol resistance. J. Bacteriol. *150,* 815-825.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. Science *327,* 167-170.

Hsiao, W.W.L., Ung, K., Aeschliman, D., Bryan, J., Finlay, B.B., and Brinkman, F.S.L. (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. PLoS Genetics *1,* e62.

Hsiao, W., Wan, I., Jones, S.J., and Brinkman, F.S. (2003). IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics *19,* 418-420.

Hudson, C.M., Lau, B.Y., and Williams, K.P. (2015). Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes. Nucleic Acids Res. *43,* D48-53.

Ito, T., Okuma, K., Ma, X.X., Yuzawa, H., and Hiramatsu, K. (2003). Insights on antibiotic resistance of Staphylococcus aureus from its whole genome: genomic island SCC. Drug Resistance Updates *6,* 41-52.

Jain, R., Ramineni, S., and Parekh, N. (2011). IGIPT - Integrated genomic island prediction tool. Bioinformation *7,* 307-310.

Joensen, K.G., Scheutz, F., Lund, O., Hasman, H., Kaas, R.S., Nielsen, E.M., and Aarestrup, F.M. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J. Clin. Microbiol. *52,* 1501-1510.

Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J. (2014). Bacterial transformation: distribution, shared mechanisms and divergent control. Nature Reviews Microbiology *12,* 181-196.

Jones, C.S., Osborne, D.J., and Stanley, J. (1992). Enterobacterial tetracycline resistance in relation to plasmid incompatibility. Mol. Cell. Probes *6,* 313-317.

Jones, G.S., and D'Orazio, S.E. (2013). Listeria monocytogenes: cultivation and laboratory maintenance. Current Protocols in Microbiology 9B. 2.1-9B. 2.7.

Junttila, J.R., Niemelä, S., and Hirn, J. (1988). Minimum growth temperatures of Listeria monocytogenes and non-haemolytic listeria. J. Appl. Bacteriol. *65,* 321-327.

Kaatz, G.W., Seo, S.M., and Ruble, C.A. (1993). Efflux-mediated fluoroquinolone resistance in Staphylococcus aureus. Antimicrob. Agents Chemother. *37,* 1086-1094.

Kado, C.I. (2014). Historical Events That Spawned the Field of Plasmid Biology. Microbiol. Spectr. *2,* 10.1128/microbiolspec.PLAS-0019-2013.

Kai, M., Matsuoka, M., Nakata, N., Maeda, S., Gidoh, M., Maeda, Y., Hashimoto, K., Kobayashi, K., and Kashiwabara, Y. (1999). Diaminodiphenylsulfone resistance of Mycobacterium leprae due to mutations in the dihydropteroate synthase gene. FEMS Microbiol. Lett. *177,* 231-235.

Karlin, S., Mrázek, J., and Campbell, A.M. (1998). Codon usages in different gene classes of the Escherichia coli genome. Mol. Microbiol. *29,* 1341-1355.

Kehrenberg, C., Aarestrup, F.M., and Schwarz, S. (2007). IS21-558 insertion sequences are involved in the mobility of the multiresistance gene cfr. Antimicrob. Agents Chemother. *51,* 483-487.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12,* 996-1006.

Khan, S.A., and Novick, R.P. (1983). Complete nucleotide sequence of pT181, a tetracycline-resistance plasmid from Staphylococcus aureus. Plasmid *10,* 251-259.

Kim, D.H., Lees, W.J., Kempsell, K.E., Lane, W.S., Duncan, K., and Walsh, C.T. (1996). Characterization of a Cys115 to Asp substitution in the Escherichia coli cell wall biosynthetic enzyme UDP-GlcNAc enolpyruvyl transferase (MurA) that confers resistance to inactivation by the antibiotic fosfomycin. Biochemistry (N. Y.) *35,* 4923-4928.

Klemm, P., and Schembri, M.A. (2000). Bacterial adhesins: function and structure. International Journal of Medical Microbiology *290,* 27-35.

Klumpp, J., and Loessner, M.J. (2013). Listeria phages: genomes, evolution, and application. Bacteriophage *3,* e26861.

Knapp, C.W., Dolfing, J., Ehlert, P.A., and Graham, D.W. (2009). Evidence of increasing antibiotic resistance gene abundances in archived soils since 1940. Environ. Sci. Technol. *44,* 580-587.

Knox, J.R. (1995). Extended-spectrum and inhibitor-resistant TEM-type beta-lactamases: mutations, specificity, and three-dimensional structure. Antimicrob. Agents Chemother. *39,* 2593-2601.

Kobayashi, T., Uozumi, T., and Beppu, T. (1986). Cloning and characterization of the streptothricin-resistance gene which encodes streptothricin acetyltransferase from Streptomyces lavendulae. J. Antibiot. *39,* 688-693.

Kobe, B., and Kajava, A.V. (2001). The leucine-rich repeat as a protein recognition motif. Curr. Opin. Struct. Biol. *11,* 725-732.

Kocks, C., Gouin, E., Tabouret, M., Berche, P., Ohayon, H., and Cossart, P. (1992). L. monocytogenes-induced actin assembly requires the actA gene product, a surface protein. Cell *68,* 521-531.

Kocks, C., Hellio, R., Gounon, P., Ohayon, H., and Cossart, P. (1993). Polarized distribution of Listeria monocytogenes surface protein ActA at the site of directional actin assembly. J. Cell. Sci. *105 (Pt 3),* 699-710.

Komp Lindgren, P., Karlsson, A., and Hughes, D. (2003). Mutation rate and evolution of fluoroquinolone resistance in Escherichia coli isolates from patients with urinary tract infections. Antimicrob. Agents Chemother. *47,* 3222-3232.

Koren, S., and Phillippy, A.M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr. Opin. Microbiol. *23,* 110-120.

Kovacevic, J., Arguedas-Villa, C., Wozniak, A., Tasara, T., and Allen, K.J. (2013). Examination of food chain-derived Listeria monocytogenes strains of different serotypes reveals considerable diversity in inlA genotypes, mutability, and adaptation to cold temperatures. Appl. Environ. Microbiol. *79,* 1915-1922.

Kristiansen, B.E., Radstrom, P., Jenkins, A., Ask, E., Facinelli, B., and Sköld, O. (1990). Cloning and characterization of a DNA fragment that confers sulfonamide resistance in a serogroup B, serotype 15 strain of Neisseria meningitidis. Antimicrob. Agents Chemother. *34,* 2277-2279.

Krzywinski, M., Schein, J., Birol, İ, Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19,* 1639-1645.

Kuenne, C., Billion, A., Mraheil, M.A., Strittmatter, A., Daniel, R., Goesmann, A., Barbuddhe, S., Hain, T., and Chakraborty, T. (2013). Reassessment of the Listeria monocytogenes pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome. BMC Genomics *14,* 1.

Laabei, M., Recker, M., Rudkin, J.K., Aldeljawi, M., Gulay, Z., Sloan, T.J., Williams, P., Endres, J.L., Bayles, K.W., Fey, P.D*., et al.* (2014). Predicting the virulence of MRSA from its genome sequence. Genome Res. *24,* 839-849.

Laird, M.R., Langille, M.G., and Brinkman, F.S. (2015). GenomeD3Plot: a library for rich, interactive visualizations of genomic data in web applications. Bioinformatics *31,* 3348-3349.

Lambert, P.A. (2005). Bacterial resistance to antibiotics: modified target sites. Adv. Drug Deliv. Rev. *57,* 1471-1485.

Landman, D., Georgescu, C., Martin, D.A., and Quale, J. (2008). Polymyxins revisited. Clin. Microbiol. Rev. *21,* 449-465.

Langille, M., Hsiao, W., and Brinkman, F. (2008). Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics *9,* 329.

Langille, M.G.I., and Brinkman, F.S.L. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics *25,* 664.

Langille, M.G., Laird, M.R., Hsiao, W.W., Chiu, T.A., Eisen, J.A., and Brinkman, F.S. (2012). MicrobeDB: a locally maintainable database of microbial genomic sequences. Bioinformatics *28,* 1947-1948.

Lanza, V.F., Tedim, A.P., Martínez, J.L., Baquero, F., and Coque, T.M. (2015). The plasmidome of Firmicutes: impact on the emergence and the spread of resistance to antimicrobials. Microbiology Spectrum *3,*

Lawrence, J.G., and Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. *44,* 383-397.

Leclercq, R. (2002). Mechanisms of resistance to macrolides and lincosamides: nature of the resistance elements and their clinical implications. Clin. Infect. Dis. *34,* 482-492.

Leclercq, R., and Courvalin, P. (1991). Bacterial resistance to macrolide, lincosamide, and streptogramin antibiotics by target modification. Antimicrob. Agents Chemother. *35,* 1267-1272.

Ledala, N., Sengupta, M., Muthaiyan, A., Wilkinson, B.J., and Jayaswal, R.K. (2010). Transcriptomic response of Listeria monocytogenes to iron limitation and Fur mutation. Appl. Environ. Microbiol. *76,* 406-416.

Lederberg, J. (1998). Plasmid (1952–1997). Plasmid *39,* 1-9.

Lederberg, J. (1952). Cell genetics and hereditary symbiosis. Physiol. Rev. *32,* 403-430.

Lederberg, J., and Tatum, E.L. (1946). Gene recombination in Escherichia coli. Nature *158,* 558.

Lee, C., Chen, Y.P., Yao, T., Ma, C., Lo, W., Lyu, P., and Tang, C.Y. (2013). GI-POP: A combinational annotation and genomic island prediction pipeline for ongoing microbial genome projects. Gene *518,* 114-123.

Lee, H., Hsu, F.F., Turk, J., and Groisman, E.A. (2004). The PmrA-regulated pmrC gene mediates phosphoethanolamine modification of lipid A and polymyxin resistance in Salmonella enterica. J. Bacteriol. *186,* 4124-4133.

Lennox, E. (1955). Transduction of linked genetic characters of the host by bacteriophage P1. Virology *1,* 190-206.

Leverstein-van Hall, M.A., Box, A.T., Blok, H.E., Paauw, A., Fluit, A.C., and Verhoef, J. (2002). Evidence of extensive interspecies transfer of integron-mediated antimicrobial resistance genes among multidrug-resistant Enterobacteriaceae in a clinical setting. J. Infect. Dis. *186,* 49-56.

Li, H. (2011). Improving SNP discovery by base alignment quality. Bioinformatics *27,* 1157-1158.

Li, L.Y., Shoemaker, N.B., and Salyers, A.A. (1995). Location and characteristics of the transfer region of a Bacteroides conjugative transposon and regulation of transfer genes. J. Bacteriol. *177,* 4992-4999.

Li, W., Raoult, D., and Fournier, P.E. (2009). Bacterial strain typing in the genomic era. FEMS Microbiol. Rev. *33,* 892-916.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest.*2/3,*

Lina, G., Quaglia, A., Reverdy, M.E., Leclercq, R., Vandenesch, F., and Etienne, J. (1999). Distribution of genes encoding resistance to macrolides, lincosamides, and streptogramins among staphylococci. Antimicrob. Agents Chemother. *43,* 1062-1066.

Linares, J.F., Gustafsson, I., Baquero, F., and Martinez, J.L. (2006). Antibiotics as intermicrobial signaling agents instead of weapons. Proc. Natl. Acad. Sci. U. S. A. *103,* 19484-19489.

Liu, B., and Pop, M. (2009). ARDB—antibiotic resistance genes database. Nucleic Acids Res. *37,* D443-D447.

Liu, D. (2006). Identification, subtyping and virulence determination of Listeria monocytogenes, an important foodborne pathogen. J. Med. Microbiol. *55,* 645-659.

Liu, Y., Wang, Y., Walsh, T.R., Yi, L., Zhang, R., Spencer, J., Doi, Y., Tian, G., Dong, B., and Huang, X. (2015). Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. The Lancet Infectious Diseases

Loessner, M.J. (1991). Improved procedure for bacteriophage typing of Listeria strains and evaluation of new phages. Appl. Environ. Microbiol. *57,* 882-884.

Loman, N.J., Constantinidou, C., Chan, J.Z., Halachev, M., Sergeant, M., Penn, C.W., Robinson, E.R., and Pallen, M.J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nature Reviews Microbiology *10,* 599-606.

Long, K.S., Poehlsgaard, J., Kehrenberg, C., Schwarz, S., and Vester, B. (2006). The Cfr rRNA methyltransferase confers resistance to Phenicols, Lincosamides, Oxazolidinones, Pleuromutilins, and Streptogramin A antibiotics. Antimicrob. Agents Chemother. *50,* 2500-2505.

Lopez, P., Espinosa, M., Greenberg, B., and Lacks, S.A. (1987). Sulfonamide resistance in Streptococcus pneumoniae: DNA sequence of the gene encoding dihydropteroate synthase and characterization of the enzyme. J. Bacteriol. *169,* 4320-4326.

Lorenz, M.G., and Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Rev. *58,* 563-602.

Lu, B., and Leong, H.W. (2016a). Computational methods for predicting genomic islands in microbial genomes. Computational and Structural Biotechnology Journal

Lu, B., and Leong, H.W. (2016b). GI-SVM: A sensitive method for predicting genomic islands based on unannotated sequence of a single genome. Journal of Bioinformatics and Computational Biology

Lunetta, K.L., Hayward, L., Segal, J., and Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. BMC Genetics *5,* 1.

Lyautey, E., Hartmann, A., Pagotto, F., Tyler, K., Lapen, D.R., Wilkes, G., Piveteau, P., Rieu, A., Robertson, W.J., and Medeiros, D.T. (2007a). Characteristics and frequency of detection of fecal Listeria monocytogenes shed by livestock, wildlife, and humans. Can. J. Microbiol. *53,* 1158-1167.

Lyautey, E., Lapen, D.R., Wilkes, G., McCleary, K., Pagotto, F., Tyler, K., Hartmann, A., Piveteau, P., Rieu, A., Robertson, W.J*., et al.* (2007b). Distribution and characteristics of Listeria monocytogenes isolates from surface waters of the South Nation River watershed, Ontario, Canada. Appl. Environ. Microbiol. *73,* 5401-5410.

Mahillon, J., and Chandler, M. (1998). Insertion sequences. Microbiol. Mol. Biol. Rev. *62,* 725-774.

Malouin, F., and Bryan, L.E. (1986). Modification of penicillin-binding proteins as mechanisms of beta-lactam resistance. Antimicrob. Agents Chemother. *30,* 1-5.

Marquis, H., Goldfine, H., and Portnoy, D.A. (1997). Proteolytic pathways of activation and degradation of a bacterial phospholipase C during intracellular infection by Listeria monocytogenes. J. Cell Biol. *137,* 1381-1392.

Marrack, P., and Kappler, J. (1990). The Staphylococcal Enterotoxins and Their Relatives.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17,* pp. 10-12.

Martinez, J.L., Sanchez, M.B., Martinez-Solano, L., Hernandez, A., Garmendia, L., Fajardo, A., and Alvarez-Ortega, C. (2009). Functional role of bacterial multidrug efflux pumps in microbial natural ecosystems. FEMS Microbiol. Rev. *33,* 430-449.

Martínez-Martínez, L., Cano, M.E., Rodríguez-Martínez, J.M., Calvo, J., and Pascual, Á. (2014). Plasmid-mediated quinolone resistance. Expert Review of Anti-Infective Therapy

Martínez-Martínez, L., Pascual, A., and Jacoby, G.A. (1998). Quinolone resistance from a transferable plasmid. The Lancet *351,* 797-799.

Mazel, D. (2006). Integrons: agents of bacterial evolution. Nature Reviews Microbiology *4,* 608-620.

McArthur, A.G., Waglechner, N., Nizam, F., Yan, A., Azad, M.A., Baylay, A.J., Bhullar, K., Canova, M.J., De Pascale, G., Ejim, L.*, et al.* (2013). The comprehensive antibiotic resistance database. Antimicrob. Agents Chemother. *57,* 3348-3357.

McClintock, B. (1941). The Stability of Broken Ends of Chromosomes in Zea Mays. Genetics *26,* 234-282.

McDermott, J.E., Corrigan, A., Peterson, E., Oehmen, C., Niemann, G., Cambronne, E.D., Sharp, D., Adkins, J.N., Samudrala, R., and Heffron, F. (2011). Computational prediction of type III and IV secreted effectors in gram-negative bacteria. Infect. Immun. *79,* 23-32.

McLauchlin, J., Audurier, A., Frommelt, A., Gerner-Smidt, P., Jacquet, C., Loessner, M., Van Der Mee-Marquet, N., Rocourt, J., Shah, S., and Wilhelms, D. (1996). WHO study on subtyping Listeria monocytogenes: results of phage-typing. Int. J. Food Microbiol. *32,* 289-299.

McLauchlin, J., Audurier, A., and Taylor, A. (1986). The evaluation of a phage-typing system for Listeria monocytogenes for use in epidemiological studies. J. Med. Microbiol. *22,* 357-365.

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

Mei, Q., Gurunathan, S., Masur, H., and Kovacs, J.A. (1998). Failure of co-trimoxazole in Pneumocystis carinii infection and mutations in dihydropteroate synthase gene. The Lancet *351,* 1631-1632.

Meka, V.G., and Gold, H.S. (2004). Antimicrobial resistance to linezolid. Clin. Infect. Dis. *39,* 1010-1015.

Meka, V.G., Pillai, S.K., Sakoulas, G., Wennersten, C., Venkataraman, L., DeGirolami, P.C., Eliopoulos, G.M., Moellering, R.C.,Jr, and Gold, H.S. (2004). Linezolid resistance in sequential Staphylococcus aureus isolates associated with a T2500A mutation in the 23S rRNA gene and loss of a single copy of rRNA. J. Infect. Dis. *190,* 311-317.

Menard, R., Prevost, M.C., Gounon, P., Sansonetti, P., and Dehio, C. (1996). The secreted Ipa complex of Shigella flexneri promotes entry into mammalian cells. Proc. Natl. Acad. Sci. U. S. A. *93,* 1254-1258.

Mendes, R.E., Deshpande, L.M., Castanheira, M., DiPersio, J., Saubolle, M.A., and Jones, R.N. (2008). First report of cfr-mediated resistance to linezolid in human staphylococcal clinical isolates recovered in the United States. Antimicrob. Agents Chemother. *52,* 2244-2246.

Mendez, B., Tachibana, C., and Levy, S.B. (1980). Heterogeneity of tetracycline resistance determinants. Plasmid *3,* 99-108.

Miller, C.A., Qiao, Y., DiSera, T., D'Astous, B., and Marth, G.T. (2014). Bam. Iobio: a Web-based, real-time, sequence alignment file inspector. Nature Methods *11,* 1189-1189.

Miller, M.A., Pfeiffer, W., and Schwartz, T. Paper presented at Gateway Computing Environments Workshop (GCE), 2010.

Mingeot-Leclercq, M.P., Glupczynski, Y., and Tulkens, P.M. (1999). Aminoglycosides: activity and resistance. Antimicrob. Agents Chemother. *43,* 727-737.

Moken, M.C., McMurry, L.M., and Levy, S.B. (1997). Selection of multiple-antibiotic-resistant (mar) mutants of Escherichia coli by using the disinfectant pine oil: roles of the mar and acrAB loci. Antimicrob. Agents Chemother. *41,* 2770-2772.

Moore, R.A., and Hancock, R.E. (1986). Involvement of outer membrane of Pseudomonas cepacia in aminoglycoside and polymyxin resistance. Antimicrob. Agents Chemother. *30,* 923-926.

Morales, G., Picazo, J.J., Baos, E., Candel, F.J., Arribi, A., Pelaez, B., Andrade, R., de la Torre, M.A., Fereres, J., and Sanchez-Garcia, M. (2010). Resistance to linezolid is mediated by the cfr gene in the first report of an outbreak of linezolid-resistant Staphylococcus aureus. Clin. Infect. Dis. *50,* 821-825.

Mukhtar, T.A., Koteva, K.P., Hughes, D.W., and Wright, G.D. (2001). Vgb from Staphylococcus aureus inactivates streptogramin B antibiotics by an elimination mechanism not hydrolysis. Biochemistry (N. Y.) *40,* 8877-8886.

Murray, B.E., and Mederski-Samaroj, B. (1983). Transferable beta-lactamase. A new mechanism for in vitro penicillin resistance in Streptococcus faecalis. J. Clin. Invest. *72,* 1168-1171.

Naas, T., Blot, M., Fitch, W.M., and Arber, W. (1994). Insertion sequence-related genetic variation in resting Escherichia coli K-12. Genetics *136,* 721-730.

Naas, T., Sougakoff, W., Casetta, A., and Nordmann, P. (1998). Molecular characterization of OXA-20, a novel class D beta-lactamase, and its integron from Pseudomonas aeruginosa. Antimicrob. Agents Chemother. *42,* 2074-2083.

Naglich, J.G., and Andrews, R.E. (1988). Tn916-dependent conjugal transfer of pC194 and pUB110 from Bacillus subtilis into Bacillus thuringiensis subsp. israelensis. Plasmid *20,* 113-126.

Naiemi, N.A., Duim, B., Savelkoul, P.H., Spanjaard, L., de Jonge, E., Bart, A., Vandenbroucke-Grauls, C.M., and de Jong, M.D. (2005). Widespread transfer of resistance genes between bacterial species in an intensive care unit: implications for hospital epidemiology. J. Clin. Microbiol. *43,* 4862-4864.

Nelson, K.E., Fouts, D.E., Mongodin, E.F., Ravel, J., DeBoy, R.T., Kolonay, J.F., Rasko, D.A., Angiuoli, S.V., Gill, S.R., Paulsen, I.T.*, 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.

Nemec, A., Dolzani, L., Brisse, S., van den Broek, P., and Dijkshoorn, L. (2004). Diversity of aminoglycoside-resistance genes and their association with class 1 integrons among strains of pan-European Acinetobacter baumannii clones. J. Med. Microbiol. *53,* 1233-1240.

Newton, B.A. (1956). The properties and mode of action of the polymyxins. Bacteriol. Rev. *20,* 14-27.

Nguyen, S.V., and McShan, W.M. (2014). Chromosomal islands of Streptococcus pyogenes and related streptococci: molecular switches for survival and virulence. Front. Cell. Infect. Microbiol. *4,* 109.

Nightingale, K.K., Ivy, R.A., Ho, A.J., Fortes, E.D., Njaa, B.L., Peters, R.M., and Wiedmann, M. (2008). inlA premature stop codons are common among Listeria monocytogenes isolates from foods and yield virulence-attenuated strains that confer protection against fully virulent strains. Appl. Environ. Microbiol. *74,* 6570-6583.

Nightingale, K.K., Windham, K., Martin, K.E., Yeung, M., and 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.

Nikaido, H. Paper presented at Seminars in cell & developmental biology.

Normark, B.H., and Normark, S. (2002). Evolution and spread of antibiotic resistance. J. Intern. Med. *252,* 91-106.

Nufer, U., Stephan, R., and Tasara, T. (2007). Growth characteristics of Listeria monocytogenes, Listeria welshimeri and Listeria innocua strains in broth cultures and a sliced bologna-type product at 4 and 7 C. Food Microbiol. *24,* 444-451.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405,* 299-304.

Ohnishi, M., Kurokawa, K., and Hayashi, T. (2001). Diversification of Escherichia coli genomes: are bacteriophages the major contributors? Trends Microbiol. *9,* 481-485.

Olaitan, A.O., Morand, S., and Rolain, J.M. (2014). Mechanisms of polymyxin resistance: acquired and intrinsic resistance in bacteria. Front. Microbiol. *5,* 643.

Orata, F.D., Keim, P.S., and Boucher, Y. (2014). The 2010 cholera outbreak in Haiti: how science solved a controversy. PLoS Pathog *10,* e1003967.

Orsi, R.H., Borowsky, M.L., Lauer, P., Young, S.K., Nusbaum, C., Galagan, J.E., Birren, B.W., Ivy, R.A., Sun, Q., and Graves, L.M. (2008). Short-term genome evolution of Listeria monocytogenes in a non-controlled environment. BMC Genomics *9,* 1.

Ou, H.Y., He, X., Harrison, E.M., Kulasekara, B.R., Thani, A.B., Kadioglu, A., Lory, S., Hinton, J.C.D., Barer, M.R., and Deng, Z. (2007). MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands. Nucleic Acids Res. *35,* W97.

Ouellette, M., Bissonnette, L., and Roy, P.H. (1987). Precise insertion of antibiotic resistance determinants into Tn21-like transposons: nucleotide sequence of the OXA-1 beta-lactamase gene. Proc. Natl. Acad. Sci. U. S. A. *84,* 7378-7382.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics *31,* 3691-3693.

Pallen, M.J., and Wren, B.W. (2007). Bacterial pathogenomics. Nature *449,* 835-842.

Papanicolaou, G.A., Medeiros, A.A., and Jacoby, G.A. (1990). Novel plasmid-mediated beta-lactamase (MIR-1) conferring resistance to oxyimino- and alpha-methoxy beta-lactams in clinical isolates of Klebsiella pneumoniae. Antimicrob. Agents Chemother. *34,* 2200-2209.

Parkhill, J., and Wren, B.W. (2011). Bacterial epidemiology and biology-lessons from genome sequencing. Genome Biol. *12,* 230.

Parks, D.H., Porter, M., Churcher, S., Wang, S., Blouin, C., Whalley, J., Brooks, S., and Beiko, R.G. (2009). GenGIS: A geospatial information system for genomic data. Genome Res. *19,* 1896-1904.

Partridge, S.R., and Hall, R.M. (2005). Correctly identifying the streptothricin resistance gene cassette. J. Clin. Microbiol. *43,* 4298-4300.

Patel, G., Huprikar, S., Factor, S.H., Jenkins, S.G., and Calfee, D.P. (2008). Outcomes of carbapenem-resistant Klebsiella pneumoniae infection and the impact of antimicrobial and adjunctive therapies. Infection Control & Hospital Epidemiology *29,* 1099-1106.

Peirano, G., Agerso, Y., Aarestrup, F.M., and dos Prazeres Rodrigues, D. (2005). Occurrence of integrons and resistance genes among sulphonamide-resistant Shigella spp. from Brazil. J. Antimicrob. Chemother. *55,* 301-305.

Penadés, J.R., Chen, J., Quiles-Puchalt, N., Carpena, N., and Novick, R.P. (2015). Bacteriophage-mediated spread of bacterial virulence genes. Curr. Opin. Microbiol. *23,* 171-178.

Perry, J., Waglechner, N., and Wright, G. (2016). The Prehistory of Antibiotic Resistance. Cold Spring Harb Perspect. Med. *6,* 10.1101/cshperspect.a025197.

Pesci, E.C., Pearson, J.P., Seed, P.C., and Iglewski, B.H. (1997). Regulation of las and rhl quorum sensing in Pseudomonas aeruginosa. J. Bacteriol. *179,* 3127-3132.

Philippon, A., Arlet, G., and Jacoby, G.A. (2002). Plasmid-determined AmpC-type beta-lactamases. Antimicrob. Agents Chemother. *46,* 1-11.

Pightling, A.W., Petronella, N., and Pagotto, F. (2014). Choice of reference sequence and assembler for alignment of Listeria monocytogenes short-read sequence data greatly influences rates of error in SNP analyses. PloS One *9,* e104579.

Pignatelli, M. (2016). TnT: a set of libraries for visualizing trees and track-based annotations for the web. Bioinformatics

Podell, S., Gaasterland, T., and Allen, E.E. (2008). A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. BMC Bioinformatics *9,* 1.

Podell, S., and Gaasterland, T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol. *8,* R16.

Poirel, L., Liard, A., Rodriguez-Martinez, J.M., and Nordmann, P. (2005a). Vibrionaceae as a possible source of Qnr-like quinolone resistance determinants. J. Antimicrob. Chemother. *56,* 1118-1121.

Poirel, L., Rodriguez-Martinez, J.M., Mammeri, H., Liard, A., and Nordmann, P. (2005b). Origin of plasmid-mediated quinolone resistance determinant QnrA. Antimicrob. Agents Chemother. *49,* 3523-3525.

Pollack, J.R., and Neilands, J. (1970). Enterobactin, an iron transport compound from Salmonella typhimurium. Biochem. Biophys. Res. Commun. *38,* 989-992.

Poole, K. (2005). Efflux-mediated antimicrobial resistance. J. Antimicrob. Chemother. *56,* 20-51.

Prystowsky, J., Siddiqui, F., Chosay, J., Shinabarger, D.L., Millichap, J., Peterson, L.R., and Noskin, G.A. (2001). Resistance to linezolid: characterization of mutations in rRNA and comparison of their occurrences in vancomycin-resistant enterococci. Antimicrob. Agents Chemother. *45,* 2154-2156.

Pundhir, S., Vijayvargiya, H., and Kumar, A. (2008). PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes. In Silico Biology *8,* 223-234.

Qi, J., Luo, H., and Hao, B. (2004). CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. *32,* W45.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics *13,* 341-2164-13-341.

Quan, S., Venter, H., and Dabbs, E.R. (1997). Ribosylative inactivation of rifampin by Mycobacterium smegmatis is a principal contributor to its low susceptibility to this antibiotic. Antimicrob. Agents Chemother. *41,* 2456-2460.

Quiros, P., Colomer-Lluch, M., Martinez-Castillo, A., Miro, E., Argente, M., Jofre, J., Navarro, F., and Muniesa, M. (2014). Antibiotic resistance genes in the bacteriophage DNA fraction of human fecal samples. Antimicrob. Agents Chemother. *58,* 606-609.

Radstrom, P., Fermer, C., Kristiansen, B.E., Jenkins, A., Sköld, O., and Swedberg, G. (1992). Transformational exchanges in the dihydropteroate synthase gene of Neisseria meningitidis: a novel mechanism for acquisition of sulfonamide resistance. J. Bacteriol. *174,* 6386-6393.

Rådström, P., Swedberg, G., and Sköld, O. (1991). Genetic analyses of sulfonamide resistance and its dissemination in gram-negative bacteria illustrate new aspects of R plasmid evolution. Antimicrob. Agents Chemother. *35,* 1840-1848.

Rajan, I., Aravamuthan, S., and Mande, S.S. (2007). Identification of compositionally distinct regions in genomes using the centroid method. Bioinformatics *23,* 2672.

Ramirez, M.S., Traglia, G.M., Lin, D.L., Tran, T., and Tolmasky, M.E. (2014). Plasmid-mediated antibiotic resistance and virulence in gram-negatives: the Klebsiella pneumoniae paradigm. Microbiology Spectrum *2,* 1.

Raveneau, J., Geoffroy, C., Beretti, J.L., Gaillard, J.L., Alouf, J.E., and Berche, P. (1992). Reduced virulence of a Listeria monocytogenes phospholipase-deficient mutant obtained by transposon insertion into the zinc metalloprotease gene. Infect. Immun. *60,* 916-921.

Read, T.D., and Massey, R.C. (2014). Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. Genome Med *6,* 109.

Recchia, G.D., and Hall, R.M. (1995). Gene cassettes: a new class of mobile element. Microbiology *141,* 3015-3027.

Reid, C.J., Chowdhury, P.R., and Djordjevic, S.P. (2015). Tn6026 and Tn6029 are found in complex resistance regions mobilised by diverse plasmids and chromosomal islands in multiple antibiotic resistant Enterobacteriaceae. Plasmid *80,* 127-137.

Relman, D.A. (2011). Microbial Genomics and Infectious Diseases. N. Engl. J. Med. *2011,* 347-357.

Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Van Loo, P., Van Den Bossche, M., Catthoor, K., and Sabbe, B. (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nat. Biotechnol. *30,* 61-68.

Ribeiro, V., Mujahid, S., Orsi, R., Bergholz, T., Wiedmann, M., Boor, K., and Destro, M. (2014). Contributions of σ B and PrfA to Listeria monocytogenes salt stress under food relevant conditions. Int. J. Food Microbiol. *177,* 98-108.

Rissman, A.I., Mau, B., Biehl, B.S., Darling, A.E., Glasner, J.D., and Perna, N.T. (2009). Reordering contigs of draft genomes using the Mauve Aligner. Bioinformatics *25,* 2071.

Roberts, M. (1996). Tetracycline resistance determinants: mechanisms of action, regulation of expression, genetic mobility, and distribution. FEMS Microbiol. Rev. *19,* 1-24.

Rouch, D., Messerotti, L., Loo, L., Jackson, C., and Skurray, R. (1989). Trimethoprim resistance transposon Tn4003 from Staphylococcus aureus encodes genes for a dihydrofolate reductase and thymidylate synthetase flanked by three copies of IS257. Mol. Microbiol. *3,* 161-175.

Russell, N.J. (1990). Cold adaptation of microorganisms. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *326,* 595-608, discussion 608-11.

Salyers, A.A., Shoemaker, N.B., and Li, L.Y. (1995a). In the driver's seat: the Bacteroides conjugative transposons and the elements they mobilize. J. Bacteriol. *177,* 5727-5731.

Salyers, A.A., Shoemaker, N.B., Stevens, A.M., and Li, L.Y. (1995b). Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. Microbiol. Rev. *59,* 579-590.

Samudrala, R., Heffron, F., and McDermott, J.E. (2009). Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. PLoS Pathogens *5,* e1000375.

Sanders, C.C., and Sanders, W.E. (1983). Emergence of resistance during therapy with the newer β-lactam antibiotics: role of inducible β-lactamases and implications for the future. Review of Infectious Diseases *5,* 639-648.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., and Bennetzen, J.L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science *274,* 765-768.

Scallan, E., Hoekstra, R.M., Angulo, F.J., Tauxe, R.V., Widdowson, M., Roy, S.L., Jones, J.L., and Griffin, P.M. (2011). Foodborne illness acquired in the United States—major pathogens. Emerging Infect. Dis. *17,*

Schlech III, W.F., Lavigne, P.M., Bortolussi, R.A., Allen, A.C., Haldane, E.V., Wort, A.J., Hightower, A.W., Johnson, S.E., King, S.H., and Nicholls, E.S. (1983). Epidemic listeriosis—evidence for transmission by food. N. Engl. J. Med. *308,* 203-206.

Schmidt, H. (2001). Shiga-toxin-converting bacteriophages. Res. Microbiol. *152,* 687-695.

Schmidt, H., and Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. Clin. Microbiol. Rev. *17,* 14-56.

Schmutz, E., Muhlenweg, A., Li, S.M., and Heide, L. (2003). Resistance genes of aminocoumarin producers: two type II topoisomerase genes confer resistance against coumermycin A1 and clorobiocin. Antimicrob. Agents Chemother. *47,* 869-877.

Schnappinger, D., and Hillen, W. (1996). Tetracyclines: antibiotic action, uptake, and resistance mechanisms. Arch. Microbiol. *165,* 359-369.

Schurch, A.C., Kremer, K., Daviena, O., Kiers, A., Boeree, M.J., Siezen, R.J., and van Soolingen, D. (2010). High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J. Clin. Microbiol. *48,* 3403.

Schwarz, S., Cardoso, M., and Wegener, H.C. (1992). Nucleotide sequence and phylogeny of the tet(L) tetracycline resistance determinant encoded by plasmid pSTE1 from Staphylococcus hyicus. Antimicrob. Agents Chemother. *36,* 580-588.

Schwender, H., Zucknick, M., Ickstadt, K., Bolt, H.M., and GENICA network. (2004). A pilot study on the application of statistical classification procedures to molecular epidemiological data. Toxicol. Lett. *151,* 291-299.

Scott, J.R. (1992). Sex and the single circle: conjugative transposition. J. Bacteriol. *174,* 6005-6010.

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

Shakil, S., Khan, R., Zarrilli, R., and Khan, A.U. (2008). Aminoglycosides versus bacteria–a description of the action, resistance mechanism, and nosocomial battleground. J. Biomed. Sci. *15,* 5-14.

Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., and Shen, B. (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed. Res. Int. *2014,* 309650.

Shapiro, J.A., and Sporn, P. (1977). Tn402: a new transposable element determining trimethoprim resistance that inserts in bacteriophage lambda. J. Bacteriol. *129,* 1632-1635.

Shaw, W.V., and Brodsky, R.F. (1968). Characterization of chloramphenicol acetyltransferase from chloramphenicol-resistant Staphylococcus aureus. J. Bacteriol. *95,* 28-36.

Shay, J. (2016). Analysis of Genomic Islands and Other Features in Draft Versus Complete Bacterial Genomes.

Sheppard, S.K., Didelot, X., Meric, G., Torralbo, A., Jolley, K.A., Kelly, D.J., Bentley, S.D., Maiden, M.C., Parkhill, J., and Falush, D. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc. Natl. Acad. Sci. U. S. A. *110,* 11923-11927.

Shrivastava, S., Reddy, Ch V Siva Kumar, and Mande, S.S. (2010). INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. J. Biosci. *35,* 351-364.

Sidhu, M.S., Heir, E., Leegaard, T., Wiger, K., and Holck, A. (2002). Frequency of disinfectant resistance genes and genetic linkage with beta-lactamase transposon Tn552 among clinical staphylococci. Antimicrob. Agents Chemother. *46,* 2797-2803.

Siguier, P., Gourbeyre, E., Varani, A., Ton-Hoang, B., and Chandler, M. (2015). Everyman's Guide to Bacterial Insertion Sequences. Microbiology Spectrum *3,*

Singh, A., Goering, R.V., Simjee, S., Foley, S.L., and Zervos, M.J. (2006). Application of molecular techniques to the study of hospital infection. Clin. Microbiol. Rev. *19,* 512.

Singh, R., Schroeder, C.M., Meng, J., White, D.G., McDermott, P.F., Wagner, D.D., Yang, H., Simjee, S., Debroy, C., Walker, R.D., and Zhao, S. (2005). Identification of antimicrobial resistance and class 1 integrons in Shiga toxin-producing Escherichia coli recovered from humans and food animals. J. Antimicrob. Chemother. *56,* 216-219.

Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J., and Holmes, I.H. (2009). JBrowse: a next-generation genome browser. Genome Res. *19,* 1630-1638.

Sköld, O. (2000). Sulfonamide resistance: mechanisms and trends. Drug Resistance Updates *3,* 155-160.

Sköld, O. (1976). R-factor-mediated resistance to sulfonamides by a plasmid-borne, drug-resistant dihydropteroate synthase. Antimicrob. Agents Chemother. *9,* 49-54.

Skold, O., and Widh, A. (1974). A new dihydrofolate reductase with low trimethoprim sensitivity induced by an R factor mediating high resistance to trimethoprim. J. Biol. Chem. *249,* 4324-4325.

Skov, R., and Monnet, D. (2016). Plasmid-mediated colistin resistance (mcr-1 gene): three months later, the story unfolds. Euro Surveill *21,* 30155.

Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P., and de la Cruz, F. (2010). Mobility of plasmids. Microbiol. Mol. Biol. Rev. *74,* 434-452.

Smit, A.F., and Riggs, A.D. (1996). Tiggers and DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. U. S. A. *93,* 1443-1448.

Smith, G.A., Marquis, H., Jones, S., Johnston, N.C., Portnoy, D.A., and Goldfine, H. (1995). The two distinct phospholipases C of Listeria monocytogenes have overlapping roles in escape from a vacuole and cell-to-cell spread. Infect. Immun. *63,* 4231-4237.

Smith, H. (1977). Microbial surfaces in relation to pathogenicity. Bacteriol. Rev. *41,* 475-500.

Smits, S.A., and Ouverney, C.C. (2010). jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. PLoS One *5,* e12267.

Snyder, L.A., and Saunders, N.J. (2006). The majority of genes in the pathogenic Neisseria species are present in non-pathogenic Neisseria lactamica, including those designated as' virulence genes'. BMC Genomics *7,* 1.

Soares, S.C., Geyik, H., Ramos, R.T., de Sá, P.H., Barbosa, E.G., Baumbach, J., Figueiredo, H.C., Miyoshi, A., Tauch, A., and Silva, A. (2015). GIPSy: Genomic island prediction software. J. Biotechnol.

Soares, S.C., Abreu, V.A., Ramos, R.T., Cerdeira, L., Silva, A., Baumbach, J., Trost, E., Tauch, A., Hirata, R.,Jr, Mattos-Guaraldi, A.L., Miyoshi, A., and Azevedo, V. (2012). PIPS: pathogenicity island prediction software. PLoS One *7,* e30848.

Sorek, R., Kunin, V., and Hugenholtz, P. (2008). CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Reviews Microbiology *6,* 181-186.

Spanogiannopoulos, P., Thaker, M., Koteva, K., Waglechner, N., and Wright, G.D. (2012). Characterization of a rifampin-inactivating glycosyltransferase from a screen of environmental actinomycetes. Antimicrob. Agents Chemother. *56,* 5061-5069.

Speer, B.S., Bedzyk, L., and Salyers, A.A. (1991). Evidence that a novel tetracycline resistance gene found on two Bacteroides transposons encodes an NADP-requiring oxidoreductase. J. Bacteriol. *173,* 176-183.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30,* 1312-1313.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., and Lewis, S. (2002). The generic genome browser: a building block for a model organism system database. Genome Res. *12,* 1599-1610.

Stokes, H.t., and Hall, R.M. (1989). A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. Mol. Microbiol. *3,* 1669-1683.

Stokes, H.W., Elbourne, L.D., and Hall, R.M. (2007). Tn1403, a multiple-antibiotic resistance transposon made up of three distinct transposons. Antimicrob. Agents Chemother. *51,* 1827-1829.

Stromberg, N., Marklund, B.I., Lund, B., Ilver, D., Hamers, A., Gaastra, W., Karlsson, K.A., and Normark, S. (1990). Host-specificity of uropathogenic Escherichia coli depends on differences in binding specificity to Gal alpha 1-4Gal-containing isoreceptors. Embo J. *9,* 2001-2010.

Suttle, C.A. (2005). Viruses in the sea. Nature *437,* 356-361.

Swaminathan, B., and Gerner-Smidt, P. (2007). The epidemiology of human listeriosis. Microb. Infect. *9,* 1236-1243.

Swedberg, G., Ringertz, S., and Sköld, O. (1998). Sulfonamide resistance in Streptococcus pyogenes is associated with differences in the amino acid sequence of its chromosomal dihydropteroate synthase. Antimicrob. Agents Chemother. *42,* 1062-1067.

Tanaka, Y., Yazawa, K., Dabbs, E.R., Nishikawa, K., Komaki, H., Mikami, Y., Miyaji, M., Morisaki, N., and Iwasaki, S. (1996). Different rifampicin inactivation mechanisms in Nocardia and related taxa. Microbiol. Immunol. *40,* 1-4.

Taniguchi, H., Chang, B., Abe, C., Nikaido, Y., Mizuguchi, Y., and Yoshida, S.I. (1997). Molecular analysis of kanamycin and viomycin resistance in Mycobacterium smegmatis by use of the conjugation system. J. Bacteriol. *179,* 4795-4801.

Taylor, D.E., and Courvalin, P. (1988). Mechanisms of antibiotic resistance in Campylobacter species. Antimicrob. Agents Chemother. *32,* 1107-1112.

Thedieck, K., Hain, T., Mohamed, W., Tindall, B.J., Nimtz, M., Chakraborty, T., Wehland, J., and Jänsch, L. (2006). The MprF protein is required for lysinylation of phospholipids in listerial membranes and confers resistance to cationic antimicrobial peptides (CAMPs) on Listeria monocytogenes. Mol. Microbiol. *62,* 1325-1339.

Theuretzbacher, U. (2011). Resistance drives antibacterial drug development. Current Opinion in Pharmacology *11,* 433-438.

Thomas, C.M., and Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature Reviews Microbiology *3,* 711-721.

Threlfall, E.J., Ward, L.R., Ashley, A.S., and Rowe, B. (1980). Plasmid-encoded trimethoprim resistance in multiresistant epidemic Salmonella typhimurium phage types 204 and 193 in Britain. Br. Med. J. *280,* 1210-1211.

Toh, S., Xiong, L., Arias, C.A., Villegas, M.V., Lolans, K., Quinn, J., and Mankin, A.S. (2007). Acquisition of a natural resistance gene renders a clinical strain of methicillin-resistant Staphylococcus aureus resistant to the synthetic antibiotic linezolid. Mol. Microbiol. *64,* 1506-1514.

Torres, O.R., Korman, R.Z., Zahler, S.A., and Dunny, G.M. (1991). The conjugative transposon Tn925: enhancement of conjugal transfer by tetracycline in Enterococcus faecalis and mobilization of chromosomal genes in Bacillus subtilis and E. faecalis. Molecular and General Genetics MGG *225,* 395-400.

Toussaint, A., and Merlin, C. (2002). Mobile elements as a combination of functional modules. Plasmid *47,* 26-35.

Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., and van Hijum, S.A. (2013). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform *14,* 315-326.

Towner, K.J., Pearson, N.J., Pinn, P.A., and O'Grady, F. (1980). Increasing importance of plasmid-mediated trimethoprim resistance in enterobacteria: two six-month clinical surveys. Br. Med. J. *280,* 517-519.

Treangen, T.J., Ondov, B.D., Koren, S., and Phillippy, A.M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. *15,* 1-15.

Tsiodras, S., Gold, H.S., Sakoulas, G., Eliopoulos, G.M., Wennersten, C., Venkataraman, L., Moellering, R.C., and Ferraro, M.J. (2001). Linezolid resistance in a clinical isolate of Staphylococcus aureus. The Lancet *358,* 207-208.

Tu, Q., and Ding, D. (2003). Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. FEMS Microbiol. Lett. *221,* 269-275.

Vakulenko, S.B., Donabedian, S.M., Voskresenskiy, A.M., Zervos, M.J., Lerner, S.A., and Chow, J.W. (2003). Multiplex PCR for detection of aminoglycoside resistance genes in enterococci. Antimicrob. Agents Chemother. *47,* 1423-1426.

Valenzuela, J.K., Thomas, L., Partridge, S.R., van der Reijden, T., Dijkshoorn, L., and Iredell, J. (2007). Horizontal gene transfer in a polyclonal outbreak of carbapenem-resistant Acinetobacter baumannii. J. Clin. Microbiol. *45,* 453-460.

Van Bambeke, F., Balzi, E., and Tulkens, P.M. (2000). Antibiotic efflux pumps. Biochem. Pharmacol. *60,* 457-470.

Van Der Veen, S., Moezelaar, R., Abee, T., and Wells-Bennik, M.H. (2008). The growth limits of a large number of Listeria monocytogenes strains at combinations of stresses show serotype-and niche-specific traits. J. Appl. Microbiol. *105,* 1246-1258.

van Treeck, U., Schmidt, F., and Wiedemann, B. (1981). Molecular nature of a streptomycin and sulfonamide resistance plasmid (pBP1) prevalent in clinical Escherichia coli strains and integration of an ampicillin resistance transposon (TnA). Antimicrob. Agents Chemother. *19,* 371-380.

Vázquez-Boland, J.A., Domínguez-Bernal, G., González-Zorn, B., Kreft, J., and Goebel, W. (2001). Pathogenicity islands and virulence evolution in Listeria. Microb. Infect. *3,* 571-584.

Vazquez-Boland, J.A., Kuhn, M., Berche, P., Chakraborty, T., Dominguez-Bernal, G., Goebel, W., Gonzalez-Zorn, B., Wehland, J., and Kreft, J. (2001). Listeria pathogenesis and molecular virulence determinants. Clin. Microbiol. Rev. *14,* 584-640.

Vernikos, G.S., and Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics *22,* 2196.

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

Vernikos, G.S., and Parkhill, J. (2008). Resolving the structural features of genomic islands: a machine learning approach. Genome Res. *18,* 331-342.

Versalovic, J., and Lupski, J.R. (2002). Molecular detection and genotyping of pathogens: more accurate and rapid answers. Trends Microbiol. *10,* s15-s21.

Vivant, A., Garmyn, D., and Piveteau, P. (2013). Listeria monocytogenes, a down-to-earth pathogen. Front Cell Infect Microbiol *3,*

von Wintersdorff, C.J., Penders, J., van Niekerk, J.M., Mills, N.D., Majumder, S., van Alphen, L.B., Savelkoul, P.H., and Wolffs, P.F. (2016). Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. Frontiers in Microbiology *7,*

Vos, M., Hesselman, M.C., Te Beek, T.A., van Passel, M.W., and Eyre-Walker, A. (2015). Rates of lateral gene transfer in prokaryotes: high but why? Trends Microbiol. *23,* 598-605.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W., Surovcik, K., Meinicke, P., and Merkl, R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics *7,* 142.

Wagner, P.L., and Waldor, M.K. (2002). Bacteriophage control of bacterial virulence. Infect. Immun. *70,* 3985-3993.

Waldor, M.K., and Mekalanos, J.J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. Science *272,* 1910.

Walker, S., Archer, P., and Banks, J.G. (1990). Growth of Listeria monocytogenes at refrigeration temperatures. J. Appl. Bacteriol. *68,* 157-162.

Wallis, R.S., Pai, M., Menzies, D., Doherty, T.M., Walzl, G., Perkins, M.D., and Zumla, A. (2010). Biomarkers and diagnostics for tuberculosis: progress, needs, and translation into practice. The Lancet *375,* 1920-1937.

Walsh, C. (2000). Molecular mechanisms that confer antibacterial drug resistance. Nature *406,* 775-781.

Wang, G., Zhou, F., Olman, V., Li, F., and Xu, Y. (2010). Prediction of pathogenicity islands in enterohemorrhagic Escherichia coli O157: H7 using genomic barcodes. FEBS Lett. *584,* 194-198.

Wang, H., Fazekas, J., Booth, M., Liu, Q., and Che, D. (2011). An Integrative Approach for Genomic Island Prediction in Prokaryotic Genomes. Bioinformatics Research and Applications 404-415.

Wang, Q., Holmes, N., Martinez, E., Howard, P., Hill-Cawthorne, G., and Sintchenko, V. (2015). It Is Not All about Single Nucleotide Polymorphisms: Comparison of Mobile Genetic Elements and Deletions in Listeria monocytogenes Genomes Links Cases of Hospital-Acquired Listeriosis to the Environmental Source. J. Clin. Microbiol. *53,* 3492-3500.

Wassenaar, T.M., and Gaastra, W. (2001). Bacterial virulence: can we draw the line? FEMS Microbiol. Lett. *201,* 1-7.

Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R.*, et al.* (2014). PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. *42,* D581-91.

Weatherill, S. (2009). Report of the independent investigator into the 2008 listeriosis outbreak Canada).

Webber, M.A., and Piddock, L.J. (2003). The importance of efflux pumps in bacterial antibiotic resistance. J. Antimicrob. Chemother. *51,* 9-11.

Wehrli, W. (1983). Rifampin: mechanisms of action and resistance. Review of Infectious Diseases *5,* S407-S411.

Wei, W., Gao, F., Du, M.Z., Hua, H.L., Wang, J., and Guo, F.B. (2016). Zisland Explorer: detect genomic islands by combining homogeneity and heterogeneity properties. Brief Bioinform

Weis, J., and Seeliger, H.P. (1975). Incidence of Listeria monocytogenes in nature. Appl. Microbiol. *30,* 29-32.

Williams, K.P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res. *30,* 866-875.

Wilson, C.A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J. Mol. Biol. *297,* 233-249.

Wise, E.M.,Jr, and Abou-Donia, M.M. (1975). Sulfonamide resistance mechanism in Escherichia coli: R plasmids can determine sulfonamide-resistant dihydropteroate synthases. Proc. Natl. Acad. Sci. U. S. A. *72,* 2621-2625.

World Health Organization. (2015). Health in 2015: from Millenium Development Goals to Sustainable Development Goals.

Wozniak, R.A., and Waldor, M.K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nature Reviews Microbiology *8,* 552-563.

Wright, G.D. (2005). Bacterial resistance to antibiotics: enzymatic degradation and modification. Adv. Drug Deliv. Rev. *57,* 1451-1470.

Wright, G.D., and Poinar, H. (2012). Antibiotic resistance is ancient: implications for drug discovery. Trends Microbiol. *20,* 157-159.

Xiong, L., Kloss, P., Douthwaite, S., Andersen, N.M., Swaney, S., Shinabarger, D.L., and Mankin, A.S. (2000). Oxazolidinone resistance mutations in 23S rRNA of Escherichia coli reveal the central region of domain V as the primary site of drug action. J. Bacteriol. *182,* 5325-5331.

Xu, Z., and Hao, B. (2009). CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res. *37,* W174.

Yamada, T., Masuda, K., Shoji, K., and Hori, M. (1972). Analysis of ribosomes from viomycin-sensitive and -resistant strains of Mycobacterium smegmatis. J. Bacteriol. *112,* 1-6.

Yang, Y., Zhao, J., Morgan, R.L., Ma, W., and Jiang, T. (2010). Computational prediction of type III secreted proteins from gram-negative bacteria. BMC Bioinformatics *11,* 1.

Yang, Y., Jiang, X., Chai, B., Ma, L., Li, B., Zhang, A., Cole, J.R., Tiedje, J.M., and Zhang, T. (2016). ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. Bioinformatics

Yazawa, K., Mikami, Y., Maeda, A., Akao, M., Morisaki, N., and Iwasaki, S. (1993). Inactivation of rifampin by Nocardia brasiliensis. Antimicrob. Agents Chemother. *37,* 1313-1317.

Yazawa, K., Mikami, Y., Maeda, A., Morisaki, N., and Iwasaki, S. (1994). Phosphorylative inactivation of rifampicin by Nocardia otitidiscaviarum. J. Antimicrob. Chemother. *33,* 1127-1135.

Yoon, S., Hur, C., Kang, H., Kim, Y., Oh, T., and Kim, J.F. (2005). A computational approach for identifying pathogenicity islands in prokaryotic genomes. BMC Bioinformatics *6,* 1.

Yoon, S.H., Park, Y.K., and Kim, J.F. (2015). PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands. Nucleic Acids Res. *43,* D624-30.

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., and Foster, L.J. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics *26,* 1608.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F.M., and Larsen, M.V. (2012). Identification of acquired antimicrobial resistance genes. J. Antimicrob. Chemother. *67,* 2640-2644.

Zapun, A., Contreras-Martel, C., and Vernet, T. (2008). Penicillin-binding proteins and beta-lactam resistance. FEMS Microbiol. Rev. *32,* 361-385.

Zarrilli, R., Crispino, M., Bagattini, M., Barretta, E., Di Popolo, A., Triassi, M., and Villari, P. (2004). Molecular epidemiology of sequential outbreaks of Acinetobacter baumannii in an intensive care unit shows the emergence of carbapenem resistance. J. Clin. Microbiol. *42,* 946-953.

Zarrilli, R., Tripodi, M.F., Di Popolo, A., Fortunato, R., Bagattini, M., Crispino, M., Florio, A., Triassi, M., and Utili, R. (2005). Molecular epidemiology of high-level aminoglycoside-resistant enterococci isolated from patients in a university hospital in southern Italy. J. Antimicrob. Chemother. *56,* 827-835.

Zhang, Y., Ren, S., Li, H., Wang, Y., Fu, G., Yang, J., Qin, Z., Miao, Y., Wang, W., and Chen, R. (2003). Genome-based analysis of virulence genes in a non-biofilm-forming Staphylococcus epidermidis strain (ATCC 12228). Mol. Microbiol. *49,* 1577-1593.

Zhou, C.E., Smith, J., Lam, M., Zemla, A., Dyer, M.D., and Slezak, T. (2007). MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res. *35,* D391-4.

Zinder, N.D., and Lederberg, J. (1952). Genetic exchange in Salmonella. J. Bacteriol. *64,* 679-699.

Zurenko, G.E., Yagi, B.H., Schaadt, R.D., Allison, J.W., Kilburn, J.O., Glickman, S.E., Hutchinson, D.K., Barbachyn, M.R., and Brickner, S.J. (1996). In vitro activities of U-100592 and U-100766, novel oxazolidinone antibacterial agents. Antimicrob. Agents Chemother. *40,* 839-845.

Appendix A.

Less-biased genomes dataset

The following collection of chromosome and plasmid sequences was used as a less-biased sample of species from all available genome sequences that represents species with a minimum evolutionary distance of 0.05, that was calculated based on a previous study (Ciccarelli et al., 2006).

Table A1. Chromosomes for each species represented in less-biased dataset

Accession	Chromosome definition
NC_012483.1	Acidobacterium capsulatum ATCC 51196
NC_000854.2	Aeropyrum pernix K1
NC_003062.2	Agrobacterium fabrum str. C58 chromosome circular
NC_003063.2	Agrobacterium fabrum str. C58 chromosome linear
NC_000918.1	Aquifex aeolicus VF5
NC_000917.1	Archaeoglobus fulgidus DSM 4304
NC_003997.3	Bacillus anthracis str. Ames
NC_003909.8	Bacillus cereus ATCC 10987
NC_004722.1	Bacillus cereus ATCC 14579
NC_002570.2	Bacillus halodurans C-125
NC_017196.1	Bacillus subtilis subsp. natto BEST195
NC_004663.1	Bacteroides thetaiotaomicron VPI-5482
NC_005363.1	Bdellovibrio bacteriovorus HD100
NC_010816.1	Bifidobacterium longum DJO10A
NC_019382.1	Bordetella bronchiseptica 253
NC_018828.1	Bordetella parapertussis Bpp5
NC_018518.1	Bordetella pertussis 18323
NC_001318.1	Borrelia burgdorferi B31
NC_017249.1	Bradyrhizobium japonicum USDA 6
NC_012441.1	Brucella melitensis ATCC 23457 chromosome I
NC_012442.1	Brucella melitensis ATCC 23457 chromosome II
NC_010169.1	Brucella suis ATCC 23445 chromosome I
NC_010167.1	Brucella suis ATCC 23445 chromosome II
NC_004545.1	Buchnera aphidicola str. Bp (Baizongia pistaciae)
NC_004061.1	Buchnera aphidicola str. Sg (Schizaphis graminum)

Accession	Chromosome definition
NC_011834.1	Buchnera aphidicola str. Tuc7 (Acyrthosiphon pisum)
NC_005061.1	Candidatus Blochmannia floridanus
NC_008536.1	Candidatus Solibacter usitatus Ellin6076
NC_002696.2	Caulobacter crescentus CB15
NC_002620.2	Chlamydia muridarum Nigg
NC_017952.1	Chlamydia trachomatis E/SW3
NC_003361.3	Chlamydophila caviae GPIC
NC_002179.2	Chlamydophila pneumoniae AR39
NC_000922.1	Chlamydophila pneumoniae CWL029
NC_002491.1	Chlamydophila pneumoniae J138
NC_005043.1	Chlamydophila pneumoniae TW-183
NC_002932.3	Chlorobium tepidum TLS
NC_005085.1	Chromobacterium violaceum ATCC 12472
NC_015687.1	Clostridium acetobutylicum DSM 1731
NC_008262.1	Clostridium perfringens SM101
NC_008265.1	Clostridium phage phiSM101
NC_022777.1	Clostridium tetani 12124569 main
NC_016782.1	Corynebacterium diphtheriae 241
NC_004369.1	Corynebacterium efficiens YS-314
NC_003450.3	Corynebacterium glutamicum ATCC 13032
NC_009342.1	Corynebacterium glutamicum R
NC_011528.1	Coxiella burnetii CbuK_Q154
NC_002936.3	Dehalococcoides ethenogenes 195
NC_001263.1	Deinococcus radiodurans R1 chromosome 1
NC_001264.1	Deinococcus radiodurans R1 chromosome 2
NC_002937.3	Desulfovibrio vulgaris str. Hildenborough
NC_017312.1	Enterococcus faecalis 62
NC_017732.1	Enterococcus phage EF62phi
NC_008563.1	Escherichia coli APEC O1
NC_002655.2	Escherichia coli O157:H7 str. EDL933
NC_002695.1	Escherichia coli O157:H7 str. Sakai
NC_017448.1	Fibrobacter succinogenes subsp. succinogenes S85
NC_003454.1	Fusobacterium nucleatum subsp. nucleatum ATCC 25586
NC_017454.1	Geobacter sulfurreducens KN400
NC_005125.1	Gloeobacter violaceus PCC 7421

Accession	Chromosome definition
NC_002940.2	Haemophilus ducreyi 35000HP
NC_000907.1	Haemophilus influenzae Rd KW20
NC_002607.1	Halobacterium sp. NRC-1
NC_004917.1	Helicobacter hepaticus ATCC 51449
NC_019560.1	Helicobacter pylori Aklavik117
NC_000921.1	Helicobacter pylori J99
NC_013504.1	Lactobacillus johnsonii FI9785
NC_021514.1	Lactobacillus plantarum 16
NC_017486.1	Lactococcus lactis subsp. lactis CV56
NC_005823.1	Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130 chromosome I
NC_005824.1	Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130 chromosome II
NC_004342.2	Leptospira interrogans serovar Lai str. 56601 chromosome I
NC_004343.2	Leptospira interrogans serovar Lai str. 56601 chromosome II
NC_003212.1	Listeria innocua Clip11262
NC_020557.1	Listeria monocytogenes La111 complete genome.
NC_002973.6	Listeria monocytogenes serotype 4b str. F2365
NC_002678.2	Mesorhizobium loti MAFF303099
NC_000909.1	Methanocaldococcus jannaschii DSM 2661
NC_009135.1	Methanococcus maripaludis C5
NC_003551.1	Methanopyrus kandleri AV19
NC_003552.1	Methanosarcina acetivorans C2A
NC_003901.1	Methanosarcina mazei Go1
NC_000916.1	Methanothermobacter thermautotrophicus str. Delta H
NC_002944.2	Mycobacterium avium subsp. paratuberculosis K-10
NC_012207.1	Mycobacterium bovis BCG str. Tokyo 172
NC_002677.1	Mycobacterium leprae TN
NC_002755.2	Mycobacterium tuberculosis CDC1551
NC_000962.3	Mycobacterium tuberculosis H37Rv
NC_017503.1	Mycoplasma gallisepticum str. F
NC_000908.2	Mycoplasma genitalium G37
NC_006908.1	Mycoplasma mobile 163K
NC_005364.2	Mycoplasma mycoides subsp. mycoides SC str. PG1
NC_004432.1	Mycoplasma penetrans HF-2
NC_016807.1	Mycoplasma pneumoniae 309
NC_002771.1	Mycoplasma pulmonis UAB CTIP

Accession	Chromosome definition
NC_005213.1	Nanoarchaeum equitans Kin4-M
NC_003112.2	Neisseria meningitidis MC58
NC_003116.1	Neisseria meningitidis Z2491
NC_004757.1	Nitrosomonas europaea ATCC 19718
NC_003272.1	Nostoc sp. PCC 7120
NC_004193.1	Oceanobacillus iheyensis HTE831
NC_005303.2	Onion yellows phytoplasma OY-M
NC_017027.1	Pasteurella multocida subsp. multocida str. HN06
NC_006370.1	Photobacterium profundum SS9 chromosome 1
NC_006371.1	Photobacterium profundum SS9 chromosome 2
NC_005126.1	Photorhabdus luminescens subsp. laumondii TTO1
NC_010729.1	Porphyromonas gingivalis ATCC 33277
NC_008816.1	Prochlorococcus marinus str. AS9601
NC_005071.1	Prochlorococcus marinus str. MIT 9313
NC_005072.1	Prochlorococcus marinus subsp. pastoris str. CCMP1986
NC_002516.2	Pseudomonas aeruginosa PAO1
NC_002947.3	Pseudomonas putida KT2440
NC_004578.1	Pseudomonas syringae pv. tomato str. DC3000
NC_003364.1	Pyrobaculum aerophilum str. IM2
NC_000868.1	Pyrococcus abyssi GE5
NC_003413.1	Pyrococcus furiosus DSM 3638
NC_000961.1	Pyrococcus horikoshii OT3
NC_005296.1	Rhodopseudomonas palustris CGA009
NC_003103.1	Rickettsia conorii str. Malish 7
NC_020993.1	Rickettsia prowazekii str. Breinl
NC_003198.1	Salmonella enterica subsp. enterica serovar Typhi str. CT18
NC_004631.1	Salmonella enterica subsp. enterica serovar Typhi str. Ty2
NC_022569.1	Salmonella enterica subsp. enterica serovar Typhimurium str. DT104
NC_004347.2	Shewanella oneidensis MR-1
NC_017328.1	Shigella flexneri 2002017
NC_004741.1	Shigella flexneri 2a str. 2457T
NC_003047.1	Sinorhizobium meliloti 1021
NC_002758.2	Staphylococcus aureus subsp. aureus Mu50
NC_003923.1	Staphylococcus aureus subsp. aureus MW2
NC_002745.2	Staphylococcus aureus subsp. aureus N315

Accession	Chromosome definition
NC_004461.1	Staphylococcus epidermidis ATCC 12228
NC_004116.1	Streptococcus agalactiae 2603V/R
NC_004368.1	Streptococcus agalactiae NEM316
NC_018089.1	Streptococcus mutans GS-5
NC_014498.1	Streptococcus pneumoniae 670-6B
NC_003098.1	Streptococcus pneumoniae R6
NC_017040.1	Streptococcus pyogenes MGAS15252
NC_004070.1	Streptococcus pyogenes MGAS315
NC_003485.1	Streptococcus pyogenes MGAS8232
NC_004606.1	Streptococcus pyogenes SSI-1
NC_003155.4	Streptomyces avermitilis MA-4680
NC_003888.3	Streptomyces coelicolor A3(2)
NC_002754.1	Sulfolobus solfataricus P2
NC_003106.2	Sulfolobus tokodaii str. 7
NC_007604.1	Synechococcus elongatus PCC 7942
NC_005070.1	Synechococcus sp. WH 8102
NC_000911.1	Synechocystis sp. PCC 6803
NC_003869.1	Thermoanaerobacter tengcongensis MB4
NC_002578.1	Thermoplasma acidophilum DSM 1728
NC_002689.2	Thermoplasma volcanium GSS1
NC_000853.1	Thermotoga maritima MSB8
NC_005835.1	Thermus thermophilus HB27
NC_002967.9	Treponema denticola ATCC 35405
NC_016842.1	Treponema pallidum subsp. pertenue str. SamoaD
NC_004572.3	Tropheryma whipplei str. Twist
NC_004551.1	Tropheryma whipplei TW08/27
NC_010503.1	Ureaplasma parvum serovar 3 str. ATCC 27815
NC_002505.1	Vibrio cholerae O1 biovar El Tor str. N16961 chromosome I
NC_002506.1	Vibrio cholerae O1 biovar El Tor str. N16961 chromosome II
NC_019955.1	Vibrio parahaemolyticus BB22OP chromosome 1
NC_019971.1	Vibrio parahaemolyticus BB22OP chromosome 2
NC_005139.1	Vibrio vulnificus YJ016 chromosome I
NC_005140.1	Vibrio vulnificus YJ016 chromosome II
NC_004344.2	Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis
NC_002978.6	Wolbachia endosymbiont of Drosophila melanogaster

Accession	Chromosome definition
NC_005090.1	Wolinella succinogenes DSM 1740
NC_003919.1	Xanthomonas axonopodis pv. citri str. 306
NC_003902.1	Xanthomonas campestris pv. campestris str. ATCC 33913
NC_002488.3	Xylella fastidiosa 9a5c
NC_004556.1	Xylella fastidiosa Temecula1
NC_010159.1	Yersinia pestis Angola
NC_005810.1	Yersinia pestis biovar Microtus str. 91001
NC_004088.1	Yersinia pestis KIM10+

Table A2. Plasmids used in less-biased dataset

Accession Number	Plasmid name
NC_003064.2	Agrobacterium fabrum str. C58 plasmid At
NC_003065.3	Agrobacterium fabrum str. C58 plasmid Ti
NC_001880.1	Aquifex aeolicus VF5 plasmid ece1
NC_005707.1	Bacillus cereus ATCC 10987 plasmid pBc10987
NC_004721.2	Bacillus cereus ATCC 14579 plasmid pBClin15
NC_017194.1	Bacillus subtilis subsp. natto BEST195 plasmid pBEST195S
NC_004703.1	Bacteroides thetaiotaomicron VPI-5482 plasmid p5482
NC_004252.1	Bifidobacterium longum DJO10A plasmid pDOJH10L
NC_004253.1	Bifidobacterium longum DJO10A plasmid pDOJH10S
NC_018830.1	Bordetella parapertussis Bpp5 plasmid BPP5P1
NC_000948.1	Borrelia burgdorferi B31 plasmid cp32-1
NC_000949.1	Borrelia burgdorferi B31 plasmid cp32-3
NC_000950.1	Borrelia burgdorferi B31 plasmid cp32-4
NC_000951.1	Borrelia burgdorferi B31 plasmid cp32-6
NC_000952.1	Borrelia burgdorferi B31 plasmid cp32-7
NC_000953.1	Borrelia burgdorferi B31 plasmid cp32-8
NC_000954.1	Borrelia burgdorferi B31 plasmid cp32-9
NC_000955.2	Borrelia burgdorferi B31 plasmid lp21
NC_000956.1	Borrelia burgdorferi B31 plasmid lp56
NC_000957.1	Borrelia burgdorferi B31 plasmid lp5
NC_001849.2	Borrelia burgdorferi B31 plasmid lp17
NC_001850.1	Borrelia burgdorferi B31 plasmid lp25

Accession Number	Plasmid name
NC_001851.2	Borrelia burgdorferi B31 plasmid lp28-1
NC_001852.1	Borrelia burgdorferi B31 plasmid lp28-2
NC_001853.1	Borrelia burgdorferi B31 plasmid lp28-3
NC_001854.1	Borrelia burgdorferi B31 plasmid lp28-4
NC_001855.1	Borrelia burgdorferi B31 plasmid lp36
NC_001856.1	Borrelia burgdorferi B31 plasmid lp38
NC_001857.2	Borrelia burgdorferi B31 plasmid lp54
NC_001903.1	Borrelia burgdorferi B31 plasmid cp26
NC_001904.1	Borrelia burgdorferi B31 plasmid cp9
NC_004555.1	Buchnera aphidicola str. Bp (Baizongia pistaciae) plasmid pBBp1
NC_022354.1	Campylobacter jejuni subsp. jejuni 00-2544 plasmid
NC_002182.1	Chlamydia muridarum Nigg plasmid pMoPn
NC_012631.1	Chlamydia trachomatis plasmid pSW3
NC_004720.1	Chlamydophila caviae GPIC plasmid pCpGP1
NC_015686.1	Clostridium acetobutylicum DSM 1731 plasmid pSMBa
NC_015688.1	Clostridium acetobutylicum DSM 1731 plasmid pSMBb
NC_022778.1	Clostridium tetani 12124569, plasmid p12124569
NC_004319.1	Corynebacterium efficiens YS-314 plasmid pCE2
NC_004320.1	Corynebacterium efficiens YS-314 plasmid pCE3
NC_009343.1	Corynebacterium glutamicum R plasmid pCGR1
NC_011526.1	Coxiella burnetii CbuK_Q154 plasmid pQpRS_K_Q154
NC_000958.1	Deinococcus radiodurans R1 plasmid MP1
NC_000959.1	Deinococcus radiodurans R1 plasmid CP1
NC_005863.1	Desulfovibrio vulgaris str. Hildenborough plasmid pDV
NC_009837.1	Escherichia coli APEC O1 plasmid pAPEC-O1-ColBM
NC_009838.1	Escherichia coli APEC O1 plasmid pAPEC-O1-R
NC_007414.1	Escherichia coli O157:H7 EDL933 plasmid pO157
NC_001869.1	Halobacterium sp. NRC-1 plasmid pNRC100
NC_002608.1	Halobacterium sp. NRC-1 plasmid pNRC200
NC_019561.1	Helicobacter pylori Aklavik117 plasmid p1HPAKL117
NC_019562.1	Helicobacter pylori Aklavik117 plasmid p2HPAKL117
NC_012552.1	Lactobacillus johnsonii FI9785 plasmid p9785S
NC_013505.1	Lactobacillus johnsonii FI9785 plasmid p9785L
NC_021515.1	Lactobacillus plantarum 16 plasmid Lp16A
NC_021516.1	Lactobacillus plantarum 16 plasmid Lp16C

Accession Number	Plasmid name
NC_021517.1	Lactobacillus plantarum 16 plasmid Lp16E
NC_021518.1	Lactobacillus plantarum 16 plasmid Lp16F
NC_021519.1	Lactobacillus plantarum 16 plasmid Lp16H
NC_021520.1	Lactobacillus plantarum 16 plasmid Lp16L
NC_021525.1	Lactobacillus plantarum 16 plasmid Lp16B
NC_021526.1	Lactobacillus plantarum 16 plasmid Lp16D
NC_021527.1	Lactobacillus plantarum 16 plasmid Lp16G
NC_021528.1	Lactobacillus plantarum 16 plasmid Lp16I
NC_017483.1	Lactococcus lactis subsp. lactis CV56 plasmid pCV56A
NC_017484.1	Lactococcus lactis subsp. lactis CV56 plasmid pCV56C
NC_017485.1	Lactococcus lactis subsp. lactis CV56 plasmid pCV56D
NC_017487.1	Lactococcus lactis subsp. lactis CV56 plasmid pCV56B
NC_017488.1	Lactococcus lactis subsp. lactis CV56 plasmid pCV56E
NC_003383.1	Listeria innocua Clip11262 plasmid pLI100
NC_002679.1	Mesorhizobium loti MAFF303099 plasmid pMLa
NC_002682.1	Mesorhizobium loti MAFF303099 plasmid pMLb
NC_001732.1	Methanocaldococcus jannaschii DSM 2661 plasmid large ECE
NC_001733.1	Methanocaldococcus jannaschii DSM 2661 plasmid small ECE
NC_009136.1	Methanococcus maripaludis C5 plasmid pMMC501
NC_003240.1	Nostoc sp. PCC 7120 plasmid pCC7120beta
NC_003241.1	Nostoc sp. PCC 7120 plasmid pCC7120zeta
NC_003267.1	Nostoc sp. PCC 7120 plasmid pCC7120gamma
NC_003270.1	Nostoc sp. PCC 7120 plasmid pCC7120epsilon
NC_003273.1	Nostoc sp. PCC 7120 plasmid pCC7120delta
NC_003276.1	Nostoc sp. PCC 7120 plasmid pCC7120alpha
NC_017035.1	Pasteurella multocida subsp. multocida str. HN06 plasmid unnamed
NC_005871.1	Photobacterium profundum SS9 plasmid pPBPR1
NC_004632.1	Pseudomonas syringae pv. tomato str. DC3000 plasmid pDC3000B
NC_004633.1	Pseudomonas syringae pv. tomato str. DC3000 plasmid pDC3000A
NC_001773.1	Pyrococcus abyssi GE5 plasmid pGT5
NC_001399.1	Ralstonia solanacearum M4S plasmid pJTPS1
NC_005297.1	Rhodopseudomonas palustris CGA009 plasmid pRPA
NC_003384.1	Salmonella enterica subsp. enterica serovar Typhi str. CT18 plasmid pHCM1
NC_003385.1	Salmonella enterica subsp. enterica serovar Typhi str. CT18 plasmid pHCM2

Accession Number	Plasmid name
NC_022570.1	Salmonella enterica subsp. enterica serovar Typhimurium DT104 plasmid pDT104, complete genome.
NC_004349.1	Shewanella oneidensis MR-1 plasmid megaplasmid
NC_017319.1	Shigella flexneri 2002017 plasmid pSFxv_1
NC_017320.1	Shigella flexneri 2002017 plasmid pSFxv_2
NC_017321.1	Shigella flexneri 2002017 plasmid pSFxv_4
NC_017329.1	Shigella flexneri 2002017 plasmid pSFxv_3
NC_017330.1	Shigella flexneri 2002017 plasmid pSFxv_5
NC_003037.1	Sinorhizobium meliloti 1021 plasmid pSymA
NC_003078.1	Sinorhizobium meliloti 1021 plasmid pSymB
NC_002774.1	Staphylococcus aureus subsp. aureus Mu50 plasmid VRSAp
NC_003140.1	Staphylococcus aureus subsp. aureus N315 plasmid pN315
NC_005003.1	Staphylococcus epidermidis ATCC 12228 plasmid pSE-12228-06
NC_005004.1	Staphylococcus epidermidis ATCC 12228 plasmid pSE-12228-05
NC_005005.1	Staphylococcus epidermidis ATCC 12228 plasmid pSE-12228-04
NC_005006.1	Staphylococcus epidermidis ATCC 12228 plasmid pSE-12228-03
NC_005007.1	Staphylococcus epidermidis ATCC 12228 plasmid pSE-12228-02
NC_005008.1	Staphylococcus epidermidis ATCC 12228 plasmid pSE-12228-01
NC_004719.1	Streptomyces avermitilis MA-4680 plasmid SAP1
NC_007595.1	Synechococcus elongatus PCC 7942 plasmid 1
NC_005229.1	Synechocystis sp. PCC 6803 plasmid pSYSM
NC_005230.1	Synechocystis sp. PCC 6803 plasmid pSYSA
NC_005231.1	Synechocystis sp. PCC 6803 plasmid pSYSG
NC_005232.1	Synechocystis sp. PCC 6803 plasmid pSYSX
NC_005838.1	Thermus thermophilus HB27 plasmid pTT27
NC_005128.1	Vibrio vulnificus YJ016 plasmid pYJ016
NC_003425.1	Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis plasmid pWb1
NC_003921.3	Xanthomonas axonopodis pv. citri str. 306 plasmid pXAC33
NC_003922.1	Xanthomonas axonopodis pv. citri str. 306 plasmid pXAC64
NC_002489.3	Xylella fastidiosa 9a5c plasmid pXF1.3
NC_002490.1	Xylella fastidiosa 9a5c plasmid pXF51
NC_004554.1	Xylella fastidiosa Temecula1 plasmid pXFPD1.3
NC_003903.1	Streptomyces coelicolor A3(2) plasmid SCP1
NC_003904.1	Streptomyces coelicolor A3(2) plasmid SCP2
NC_010157.1	Yersinia pestis Angola plasmid new_pCD

Accession Number	Plasmid name
NC_010158.1	Yersinia pestis Angola plasmid pMT-pPCP
NC_004838.1	Yersinia pestis KIM plasmid pMT-1
NC_005813.1	Yersinia pestis biovar Microtus str. 91001 plasmid pCD1
NC_005814.1	Yersinia pestis biovar Microtus str. 91001 plasmid pCRY
NC_005815.1	Yersinia pestis biovar Microtus str. 91001 plasmid pMT1
NC_005816.1	Yersinia pestis biovar Microtus str. 91001 plasmid pPCP1
NC_002127.1	Escherichia coli O157:H7 str. Sakai plasmid pOSAK1
NC_002128.1	Escherichia coli O157:H7 str. Sakai plasmid pO157
NC_017313.1	Enterococcus faecalis 62 plasmid EF62pB
NC_017314.1	Enterococcus faecalis 62 plasmid EF62pA
NC_017315.1	Enterococcus faecalis 62 plasmid EF62pC
NC_008263.1	Clostridium perfringens SM101 plasmid pSM101A
NC_008264.1	Clostridium perfringens SM101 plasmid pSM101B

Appendix B.

Additional data for *Listeria monocytogenes* isolates used in cold growth analysis

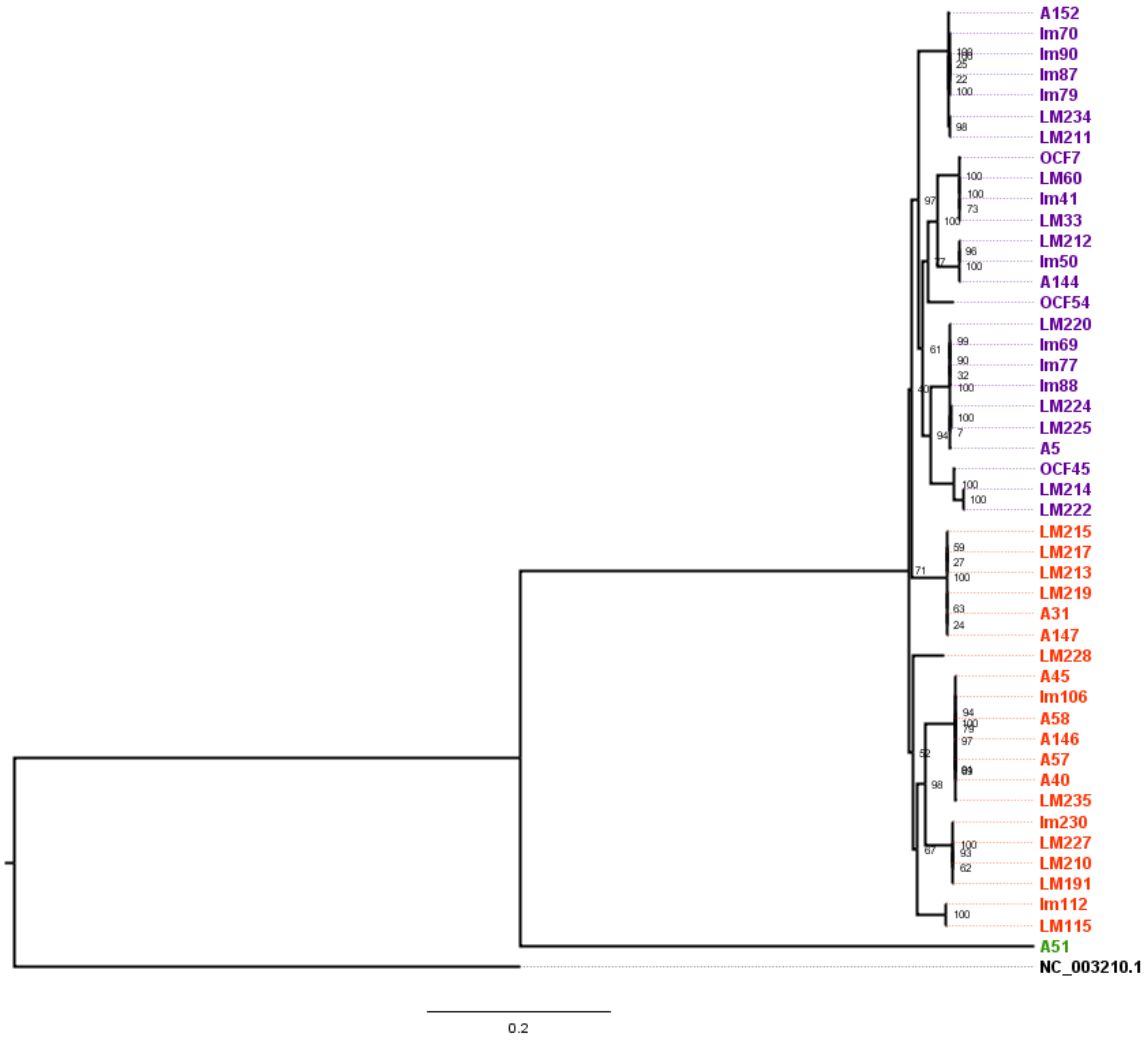

Figure B1 **Phylogenetic tree of *L. monocytogenes* isolates from serovars 1/2b (orange), 4b (purple), and 4c (green) generated from core genome SNVs as calculated by Parsnp**

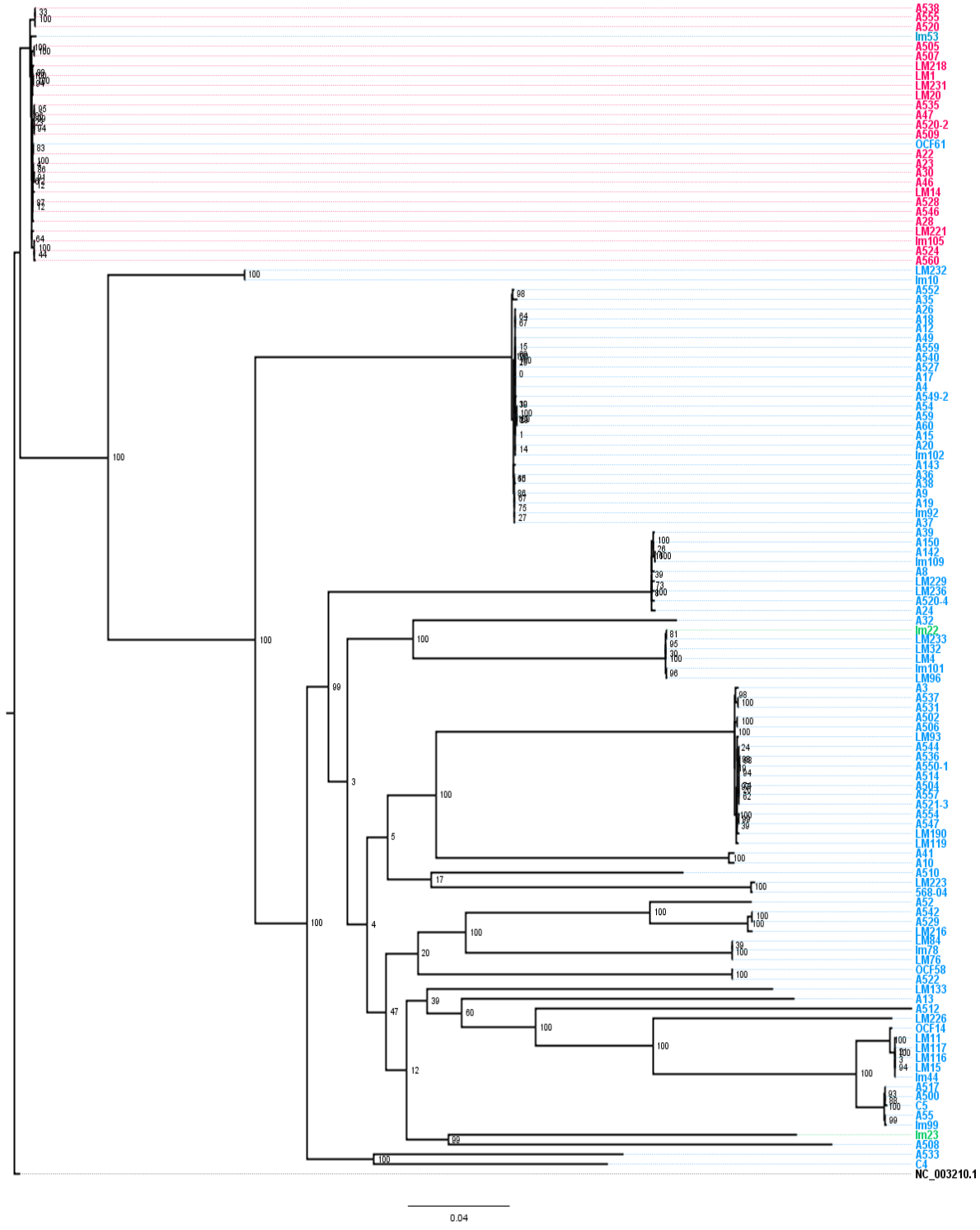

Figure B2 **Phylogenetic tree of *L. monocytogenes* isolates from serovars 1/2a (blue), 1/2c (pink), and 3a (green) generated from core genome SNVs as calculated by Parsnp**

Table B1 Select metadata associated with *L. monocytogenes* collection

Isolate	Province or Country	Sampling Date	Source	Serotype	Cold tolerance phenotype*	Average coverage	SNVs against NC_003210.2	Contigs in *de-novo* assembly
A10	AB	1/24/1990	Food	1/2a	CT	515	28,952	15
A12	AB	1/25/1990	Food	1/2a		539	24,523	18
A13	AB	1/25/1990	Food	1/2a		506	28,078	22
A142	AB	1/3/1992	Food	1/2a		586	30,885	20
A143	AB	1/3/1992	Food	1/2a		572	24,657	18
A144	AB	1/3/1992	Food	4b		490	136,440	41
A146	AB	1/3/1992	Food	1/2b		521	134,455	27
A147	AB	1/3/1992	Food	1/2b		481	132,781	285
A15	AB	1/25/1990	Food	1/2a	CT	539	24,541	20
A150	AB	1/3/1992	Food	1/2a		561	30,561	33
A152	AB	1/3/1992	Food	4b		443	135,742	29
A17	AB	1/26/1990	Food	1/2a		467	24,547	20
A18	AB	1/26/1990	Food	1/2a		442	24,542	20
A19	AB	1/26/1990	Food	1/2a		523	24,609	25
A20	AB	1/30/1990	Food	1/2a		555	24,625	20
A22	AB	1/30/1990	Food	1/2c		541	2,374	23
A23	AB	1/30/1990	Food	1/2c		494	2,285	19
A24	AB	1/31/1990	Food	1/2a		439	30,061	18
A26	AB	1/31/1990	Food	1/2a		517	24,588	20
A28	AB	1/31/1990	Food	1/2c		561	1,683	18
A3	AB	8/2/1990	Food	1/2a		585	27,835	20

Isolate	Province or Country	Sampling Date	Source	Serotype	Cold tolerance phenotype*	Average coverage	SNVs against NC_003210.2	Contigs in *de-novo* assembly
A30	AB	1/31/1990	Food	1/2c	CS	567	4,120	23
A31	AB	1/31/1990	Food	1/2b		493	133,387	25
A32	AB	1/31/1990	Food	1/2a		502	25,430	14
A35	AB	2/23/1990	Food	1/2a		467	25,051	22
A36	AB	2/23/1990	Food	1/2a		457	24,465	56
A37	AB	2/23/1990	Food	1/2a		463	24,355	77
A38	AB	2/28/1990	Food	1/2a		476	24,433	47
A39	AB	2/28/1990	Food	1/2a	CS	530	30,644	38
A4	AB	8/15/1990	Food	1/2a		523	24,606	19
A40	AB	2/28/1990	Food	1/2b	VCT	516	135,051	21
A41	AB	2/28/1990	Food	1/2a		448	31,573	14
A45	AB	3/8/1990	Food	1/2b		481	134,009	22
A46	AB	3/8/1990	Food	1/2c		534	2,003	27
A47	AB	3/8/1990	Food	1/2c		568	1,956	24
A49	AB	3/8/1990	Food	1/2a		471	24,582	57
A5	AB	8/2/1990	Food	4b	VCS	476	133,756	19
A500	AB	12/28/1990	Food	1/2a		472	34,597	25
A502	AB	12/28/1990	Food	1/2a		487	28,637	14
A504	AB	12/28/1990	Food	1/2a		491	29,462	21
A505	AB	12/28/1990	Food	1/2c	CS	465	3,901	35
A506	AB	12/28/1990	Food	1/2a		527	28,855	14
A507	AB	12/28/1990	Food	1/2c	CT	509	3,958	36
A508	AB	12/28/1990	Food	1/2a		401	32,638	19
A509	AB	12/28/1990	Food	1/2c		469	3,903	39

Isolate	Province or Country	Sampling Date	Source	Serotype	Cold tolerance phenotype*	Average coverage	SNVs against NC_003210.2	Contigs in *de-novo* assembly
A51	AB	3/8/1990	Food	4c		418	159,824	14
A510	AB	12/28/1990	Food	1/2a		572	28,444	14
A512	AB	12/28/1990	Food	1/2a		452	37,770	20
A514	AB	12/28/1990	Food	1/2a		492	29,571	22
A517	AB	12/28/1990	Food	1/2a		457	34,592	25
A52	AB	3/8/1990	Food	1/2a		517	25,527	12
A520	AB	1/2/1991	Food	1/2c		510	3,941	14
A520-2	AB	12/28/1990	Food	1/2c		466	3,920	15
A520-4	AB	12/28/1990	Food	1/2a		495	30,691	32
A521-3	AB	1/2/1991	Food	1/2a		409	29,378	20
A522	AB	1/2/1991	Food	1/2a	CT	492	27,541	18
A524	AB	1/2/1991	Food	1/2c		389	2,454	19
A527	AB	1/4/1991	Food	1/2a	CT	552	24,640	18
A528	AB	1/4/1991	Food	1/2c		472	2,295	28
A529	AB	1/4/1991	Food	1/2a		530	25,922	18
A531	AB	1/4/1991	Food	1/2a		504	28,860	36
A533	AB	1/4/1991	Food	1/2a		475	29,432	24
A535	AB	1/4/1991	Food	1/2c		537	3,983	19
A536	AB	1/4/1991	Food	1/2a		531	29,284	19
A537	AB	1/4/1991	Food	1/2a	CS	490	28,659	33
A538	AB	1/4/1991	Food	1/2c	CS	542	3,979	17
A54	AB	3/15/1990	Food	1/2a		486	24,483	17
A540	AB	1/4/1991	Food	1/2a		443	24,589	23
A542	AB	1/4/1991	Food	1/2a		511	25,882	18

Isolate	Province or Country	Sampling Date	Source	Serotype	Cold tolerance phenotype*	Average coverage	SNVs against NC_003210.2	Contigs in *de-novo* assembly
A544	AB	1/4/1991	Food	1/2a		534	29,381	22
A546	AB	1/4/1991	Food	1/2c		480	2,450	19
A547	AB	1/4/1991	Food	1/2a		462	28,786	14
A549-2	AB	1/4/1991	Food	1/2a	CT	392	24,508	20
A55	AB	3/15/1990	Food	1/2a		505	34,636	20
A550-1	AB	1/4/1991	Food	1/2a		536	29,555	14
A552	AB	1/4/1991	Food	1/2a		452	25,080	25
A554	AB	1/4/1991	Food	1/2a		570	30,053	31
A555	AB	1/4/1991	Food	1/2c		532	4,068	26
A557	AB	1/4/1991	Food	1/2a		506	29,452	20
A559	AB	1/4/1991	Food	1/2a		444	24,538	20
A560	AB	1/4/1991	Food	1/2c	CT	618	2,483	20
A57	AB	3/15/1990	Food	1/2b		519	135,731	22
A58	AB	3/15/1990	Food	1/2b		433	135,078	32
A59	AB	5/4/1990	Food	1/2a		483	24,519	21
A60	AB	5/4/1990	Food	1/2a		567	24,714	21
A8	AB	1/24/1990	Food	1/2a		575	29,979	17
A9	AB	1/24/1990	Food	1/2a		479	24,500	23
C4	NS	5/8/2012	Water	1/2a		571	26,473	16
C5	NS	3/6/2012	Water	1/2a		527	32,202	17
Lm1	BC	Aug. - Oct. 2009	Environmental	1/2a	CT	293	27,610	23
Lm10	BC	Aug. - Oct. 2009	Environmental	1/2a		532	30,852	22
Lm101	BC	Aug. - Oct. 2009	Environmental	1/2a		556	34,738	24
Lm102	BC	Aug. - Oct. 2009	Food	1/2a		473	25,123	14

Isolate	Province or Country	Sampling Date	Source	Serotype	Cold tolerance phenotype*	Average coverage	SNVs against NC_003210.2	Contigs in *de-novo* assembly
Lm105	BC	Aug. - Oct. 2009	Food	1/2c		669	3,365	14
Lm106	BC	Aug. - Oct. 2009	Environmental	1/2b		596	132,543	34
Lm109	BC	Sept. Oct. 2010	Food	1/2a		594	24,724	19
Lm11	BC	Aug. - Oct. 2009	Environmental	1/2a		431	29,304	18
Lm112	BC	Sept. Oct. 2010	Food	1/2b		571	134,529	31
Lm115	BC	7/4/1905		1/2b		437	134,125	31
Lm116	BC	7/4/1905		1/2a		419	13,046	22
Lm117	BC	6/29/1905	Environmental	1/2a		505	33,959	16
Lm119	BC	7/2/1905	Environmental	1/2a	CT	398	34,153	42
Lm133	BC	7/4/1905		1/2a		355	34,227	45
Lm14	BC	Aug. - Oct. 2009	Food	1/2c		431	2,347	27
Lm15	BC	Aug. - Oct. 2009	Food	1/2a		402	29,663	26
LM190	BC	7/4/1905	Environmental	1/2a		435	33,597	35
Lm191	BC	7/4/1905	Food	1/2b		461	133,555	24
Lm20	BC	Aug. - Oct. 2009	Environmental	1/2c		456	3,646	27
Lm210	CH	2005	Listeriosis Case	1/2b		437	132,284	24
Lm211	CH	2006	Listeriosis Case	4b		394	132,853	25
Lm212	CH	2006	Listeriosis Case	4b	CS	357	132,355	23
Lm213	CH	1999	Meat	1/2b		438	132,504	24
Lm214	CH	1999	Meat	4b	CT	384	132,348	24
Lm215	CH	1999	Meat	1/2b		437	131,122	29
Lm216	CH	2000	Meat	1/2a		400	1,477	40
Lm217	CH	2005	Listeriosis Case	1/2b	CT	404	132,018	23
Lm218	CH	2005	Listeriosis Case	1/2c	CT	477	2,304	29

Isolate	Province or Country	Sampling Date	Source	Serotype	Cold tolerance phenotype*	Average coverage	SNVs against NC_003210.2	Contigs in *de-novo* assembly
Lm219	CH	1999	Meat	1/2b		386	133,852	25
Lm22	BC	Aug. - Oct. 2009	Environmental	3a		564	25,278	17
Lm220	CH	1999	Meat	4b	CT	398	132,457	23
Lm221	CH	2001	Meat	1/2c		461	2,115	20
Lm222	CH	2011	Carcasses	4b		394	132,517	23
Lm223	CH	2011	Carcasses	1/2a	CT	438	25,719	11
Lm224	CH	2011	Carcasses	4b		483	134,179	29
Lm225	CH	2011	Carcasses	4b	VCS	437	133,259	25
Lm226	CH	2011	Carcasses	1/2a	CT	509	27,605	17
Lm227	CH	2011	Carcasses	1/2b	CT	469	133,464	23
Lm228	CH	2011	Carcasses	1/2b		404	131,635	28
Lm229	CH	2011	Environment	1/2a		458	36,102	33
Lm23	BC	Aug. - Oct. 2009	Environmental	3a		581	25,238	14
Lm230	CH	2011	Seafood (Tuna sandwich)	1/2b		612	136,495	18
Lm231	CH	2004	Asymptomatic Human Carriage	1/2c	VCS	425	2,171	16
Lm232	CH	2006	Listeriosis Case	1/2a		519	30,217	19
Lm233	CH	2002	Meat	1/2c	VCS	498	1,579	44
Lm234	CH	2011	Carcasses	4b	CS	562	134,364	21
Lm235	CH	2011	Environment	1/2b		520	134,427	35
Lm236	CH	2011	Environment	1/2a	CS	490	14,114	42
Lm32	BC	Aug. - Oct. 2009	Environmental	1/2a		490	30,746	28
Lm33	BC	Aug. - Oct. 2009	Environmental	4b		514	137,018	25
Lm4	BC	Aug. - Oct. 2009	Environmental	1/2a		402	33,496	39

Isolate	Province or Country	Sampling Date	Source	Serotype	Cold tolerance phenotype*	Average coverage	SNVs against NC_003210.2	Contigs in *de-novo* assembly
Lm41	BC	Aug. - Oct. 2009	Food	4b		502	136,856	24
Lm44	BC	Aug. - Oct. 2009	Environmental	1/2a		552	25,206	35
Lm50	BC	Aug. - Oct. 2009	Food	4b		486	136,369	26
Lm53	BC	Aug. - Oct. 2009	Food	1/2a		533	25,191	33
Lm60	BC	Aug. - Oct. 2009	Food	4b		517	136,594	28
Lm69	BC	Aug. - Oct. 2009	Environmental	4b		514	135,974	20
Lm70	BC	Aug. - Oct. 2009	Environmental	4b		602	136,334	25
Lm76	BC	Aug. - Oct. 2009	Food	1/2a	CT	530	2,276	24
Lm77	BC	Aug. - Oct. 2009	Food	4b		550	136,024	21
Lm78	BC	Aug. - Oct. 2009	Food	1/2a	CT	1043	28,603	18
Lm79	BC	Aug. - Oct. 2009	Food	4b		551	136,202	23
Lm84	BC	Aug. - Oct. 2009	Food	1/2a		523	28,270	15
Lm87	BC	Aug. - Oct. 2009	Food	4b		535	136,065	24
Lm88	BC	Aug. - Oct. 2009	Food	4b		564	136,117	22
Lm90	BC	Aug. - Oct. 2009	Food	4b		537	136,145	22
Lm92	BC	Aug. - Oct. 2009	Food	1/2a		641	28,392	16
Lm93	BC	Aug. - Oct. 2009	Food	1/2a		503	24,539	23
Lm96	BC	Aug. - Oct. 2009	Environmental	1/2a	VCS	575	29,162	28
Lm99	BC	Aug. - Oct. 2009	Environmental	1/2a		671	25,071	16
OCF07	BC	4/3/2013	Environmental	4b		528	135,166	27
OCF14	BC	5/6/2013	Environmental	1/2a		494	34,566	19
OCF45	BC	7/4/2013	Environmental	4b		577	136,374	27
OCF54	BC	8/6/2013	Seagull Droppings	4b		388	132,331	36
OCF58	BC	8/6/2013	Seagull Droppings	1/2a		475	33,814	18

Isolate	Province or Country	Sampling Date	Source	Serotype	Cold tolerance phenotype*	Average coverage	SNVs against NC_003210.2	Contigs in *de-novo* assembly
OCF61	BC	9/11/2013		1/2a		569	27,855	25
568-04	France		Food	1/2a		453	27,743	14

*Letters used in abbreviations are defined as follows: V=very, C=cold, T=tolerant, S=sensitive. Empty cells represent INT growers