Analysis of Genomic Islands and Other Features in Draft Versus Complete Bacterial Genomes

by

Julie Allison Shay

B.Sc., McMaster University, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Summer 2016

Approval

Name:	**Julie Allison Shay**
Degree:	**Master of Science**
Title:	***Analysis of Genomic Islands and Other Features in Draft Versus Complete Bacterial Genomes***
Examining Committee:	**Chair:** Dr. Esther Verheyen Professor

Dr. Fiona S.L. Brinkman
Senior Supervisor
Professor

Dr. David L. Baillie
Supervisor
Professor Emeritus

Dr. Lisa Craig
Supervisor
Associate Professor

Dr. Jack N. Chen
Internal Examiner
Professor

Date Defended/Approved: August 19, 2016

Abstract

Short-read DNA sequencing technologies have revolutionized bacterial genomics, but these technologies have limitations. It is easy to produce a high quality draft genome, but relatively costly and/or time consuming to complete a genome, so most bacterial genomes remain as drafts. Despite this, limitations of draft bacterial genomes for functional analysis have not been well assessed. To characterize the importance of missing and poor quality regions of draft genomes, analyses of COG categories and genes of medical importance were performed using analogous draft and complete genomes. A popular genomic island prediction tool, IslandViewer, was updated to allow draft genomes as input, and its ability to detect genomic islands in draft genomes was assessed. There are limitations to bacterial draft genome analysis, with respect to disproportionately missing certain types of genes. However, valuable information of medical interest, including virulence and antimicrobial resistance genes, can still be obtained from some draft genome datasets.

Keywords: bioinformatics; genomics; next generation sequencing; draft genomes; genomic islands; bacteria

I dedicate this thesis to my parents,
who support me even as I keep pursuing
opportunities in new, far-away places

Acknowledgements

I would like to express my sincere thanks to my supervisor, Dr. Fiona Brinkman, for her support and kindness. I would also like to thank my committee members, Dr. David Baillie and Dr. Lisa Craig, for their guidance. I would like to thank all members of the Brinkman lab, in particular my fellow members of the IslandViewer development team: Bhavjinder Dhillon, Matthew Laird, and Dr. Claire Bertelli.

I would like to thank Cystic Fibrosis Canada, Weyerhaeuser, and the Simon Fraser University Dean of Graduate Studies Office for their support of my degree.

I would like to thank the Bioinformatics and Genomics research groups at the National Microbiology Laboratory, and the Pseudomonas International Genomics Consortium, for providing draft genome data that has been essential for my research.

Lastly, I would like to thank Kevin Bushell, along with the rest of my family and friends who live in Vancouver, for making me feel at home in a new city.

Table of Contents

List of Tables

List of Figures

Glossary

AMR	Antimicrobial Resistance
bp	base pair
CARD	Comprehensive Antibiotic Resistance Database
CDS	Coding Sequence
COG	Cluster of Orthologous Groups
contig	contiguous sequence
DNA	Deoxyribonucleic Acid
dNTP	deoxy Nucleoside Triphosphate
dsDNA	double stranded DNA
ELIDA	Enzymatic Luminometric Inorganic pyrophosphate Detection Assay
FN	False Negative
FP	False Positive
GI	Genomic Island
HGAP	Hierarchical Genome-Assembly Process
HGT	Horizontal Gene Transfer
IRIDA	Integrated Rapid Infectious Disease Analysis
IS	Insertion Sequence
MCM	Mauve Contig Mover
MGE	Mobile Genetic Element
N50	In a genome assembly, contig length such that 50% of the base pairs in the assembly are contained within contigs of that length or larger
NGS	Next Generation Sequencing
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PSSM	Position-Specific Scoring Matrix
RBBH	Reciprocal Best BLAST Hit
RGI	Resistance Gene Identifier
SMRT	Single Molecule, Real Time
SNV	Single Nucleotide Variant
SRA	Sequence Read Archive, an NCBI database

TIR	Terminal Inverted Repeat
TN	True Negative
TP	True Positive
VF	Virulence Factor
WGS	Whole Genome Sequencing

Chapter 1.

Introduction

1.1. Next generation sequencing and draft genomes

1.1.1. Current state of bacterial genome sequencing

The current state of sequencing technology makes it easy to produce a high quality draft genome, but it is still relatively expensive or time consuming to close a genome (this will be discussed in more detail in section 1.1.4). As a result, the majority of bacterial genomes are only being sequenced to the draft stage (Figure 1.1). Long read sequencing technologies hold promise as an eventual solution to this problem, but for now Illumina short read sequencing (HiSeq, MiSeq) remains by far the most prevalent sequencing technology (Thayer, 2014). and for at least the next two years will remain as the primary method for public health agencies in Canada and the United States of America for bacterial whole genome sequencing (WGS; Gary van Domselaar, Public Health Agency of Canada, personal communication). Although both complete genome sequences and raw WGS libraries are growing exponentially, the number of WGS libraries is growing considerably faster (Figure 1.1).

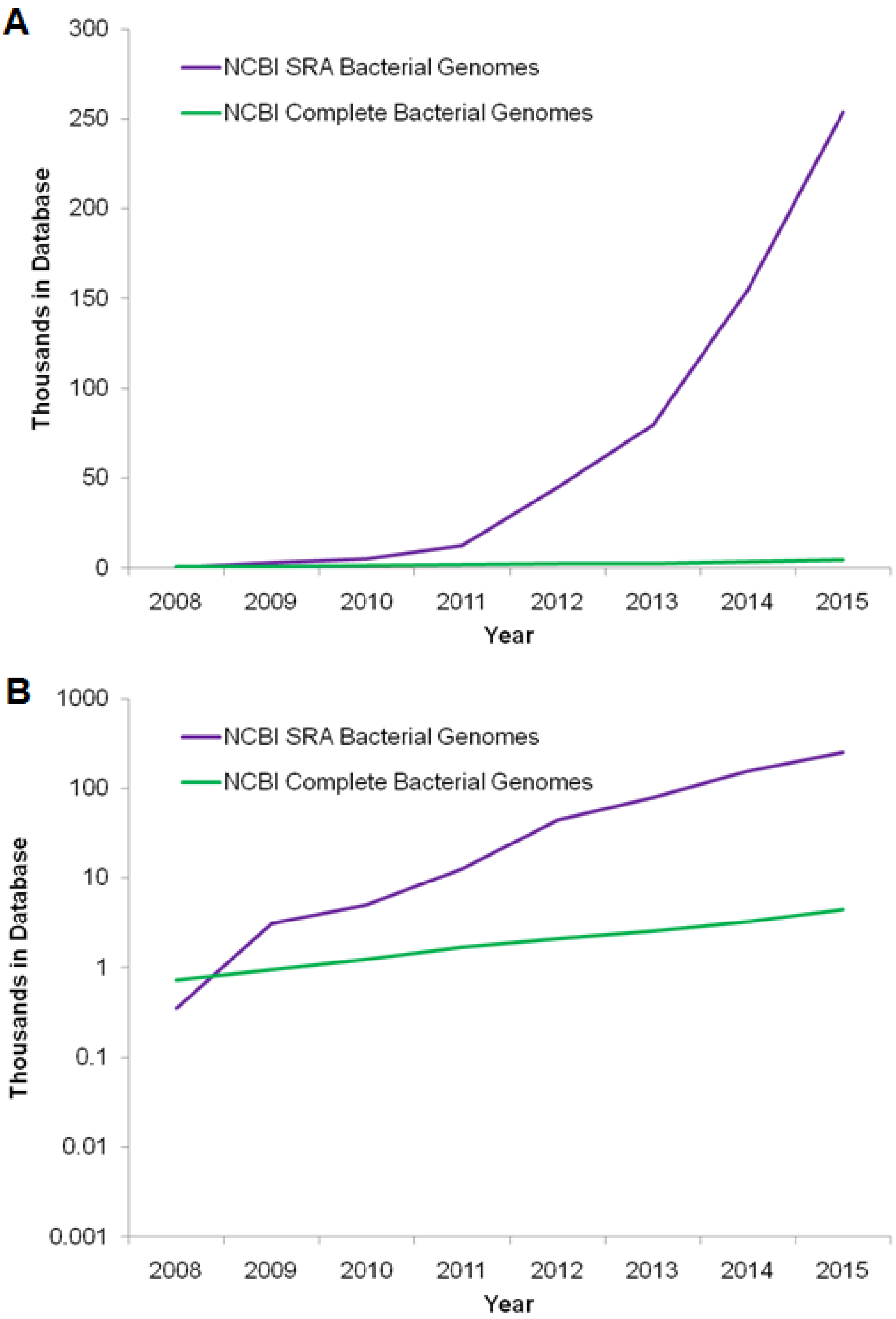


Figure 1.1 **The number of complete bacterial genomes in Genbank, and the number of bacterial genomes in the NCBI Sequence Read Archive (SRA; draft genomes). Both graphs contain the same data, but panel B has a logarithmic Y axis to emphasize the exponential growth of both curves.**

WGS of bacteria is being used increasingly in clinical and epidemiological settings (Gilchrist et al., 2015). While WGS is not yet routine in these settings, it is predicted to eventually replace traditional molecular typing methods (Gilmour et al., 2013). Canada and The United States of America are both in the process of building a network of public health labs which are capable of performing Illumina sequencing (Gary van Domselaar, National Microbiology Laboratory, personal communication, Allard et al., 2016). The US public health network alone is expected to produce thousands of bacterial genome sequences (Allard et al., 2016).

1.1.2. A brief history of DNA sequencing

Two methods to determine the sequence of adenines, guanines, cytosines, and thymines in a DNA (deoxyribonucleic acid) polymer were published in 1977. The first of these was the Maxam-Gilbert method (Maxam & Gilbert, 1977), which involves cleaving a radiolabeled DNA molecule at each instance of a certain type of base, separating the DNA fragments using electrophoresis on a polyacrylamide gel, and using the band patterns formed by cleaving these different types of bases to determine the DNA sequence. The cleavage reactions included a guanine/adenine cleavage, a preferential adenine cleavage (which also cleaved guanine), a cytosine/thymine cleavage, and a cytosine cleavage. This was the first widely adopted sequencing method (Hutchison, 2007), and could sequence about 100 base pairs (bp) of DNA.

The first paper describing dideoxy Sanger sequencing (Sanger & Nicklen, 1977) was released later that year. Frederick Sanger had developed another sequencing method earlier that was used to sequence the genome of bacteriophage ϕX174, the first complete genome (Sanger et al., 1977). This earlier method used DNA polymerase reactions that copy a template by extending a primer sequence. The first reaction produced radiolabeled copies of the original sequence with varied lengths by using radiolabeled deoxynucleoside triphosphates (dNTPs) and reaction conditions that cause the individual reactions to proceed at very different rates. These sequences were then used to prime four "plus" reactions, where only one type of nucleoside triphosphate is present per reaction, and four "minus" reactions, where all but one type of nucleoside triphosphate is present per reaction. The reaction products were then separated using electrophoresis, and the band lengths were used to determine the original sequence. The "plus and minus" method required both the plus and minus reactions because neither is sufficiently accurate on its own to determine the correct sequence (Sanger & Nicklen, 1977). The Sanger dideoxy sequencing method improved on the plus and minus method by adding a chain-terminating 2',3'-dideoxynucleoside triphosphate to the polymerase reaction. This allowed for only one polymerization reaction step (done four times, once for each nucleotide), and increased the accuracy of the resulting sequence (Sanger & Nicklen, 1977).

Changes to dideoxy Sanger sequencing have since allowed for longer sequences and more automation (Metzker, 2005). Modern Sanger sequencing uses a different fluorophore for each base (A, T, C, or G) and may either use four different reactions and label the primer sequence, or use a single reaction and label the terminating 2’,3’-dideoxynucleoside triphosphates. Rather than traditional electrophoresis, automated Sanger sequencing uses capillary array electrophoresis, which allows a computer to analyze the gel directly (Smith et al., 1986). During the human genome sequencing project, Sanger sequencing produced 500-600bp reads, and a single machine could produce 115 kbp of sequence per day (Mardis, 2011). Sanger sequencing technology was used to produce many early genome sequences, including the human genome (Lander et al., 2001; Venter et al., 2001) and several bacterial genomes (Fleischmann et al., 1995; Fraser et al., 1995; Stover et al., 2000). While it still has uses today, next generation sequencing (NGS) technologies have made Sanger sequencing obsolete for the bulk of sequencing involved in genome sequencing projects, including bacterial genome sequencing.

1.1.3. Short read sequencing technologies

Short read sequencing technologies have shorter read lengths than Sanger sequencing, but they have replaced Sanger sequencing for genome-scale sequencing projects due to their much higher throughput. The first NGS system to be released was a pyrosequencing method developed by Roche/454 (Margulies et al., 2005). Pyrosequencing gets its name from the enzymatic luminometric inorganic pyrophosphate detection assay (ELIDA), which uses sulfurylase and luciferase to fluoresce in the presence of the pyrophosphate released when a base is added to a nucleotide sequence by a polymerase (Nyrén et al., 1993). Roche/454 sequencing uses ELIDA on a very small scale: DNA is fragmented and bound to beads (approximately one fragment per bead), polymerase chain reaction (PCR) is performed on the fragments while they are attached to the bead, and then the beads containing millions of copies of the fragment template are put into extremely small wells where they are exposed to one dNTP at a time, and ELIDA is used to determine when a nucleotide is added to the new strand. A major source of error with this technology is the difficulty in determining the number of bases in a homopolymer, as this has to be measured based on the strength of signal from the ELIDA assay. According to Roche/454, their most recent machine produces 700 Mbp per run with a mode read length of 700bp (Roche Applied Science, 2011). While at one point Roche/454 sequencing was fairly popular, it is no longer in development, and its production will be shut down completely this year. Illumina technology, which I will discuss in the next paragraph, is now the most popular short read sequencing technology, taking up 71% of the world market in 2013 (Thayer, 2014).

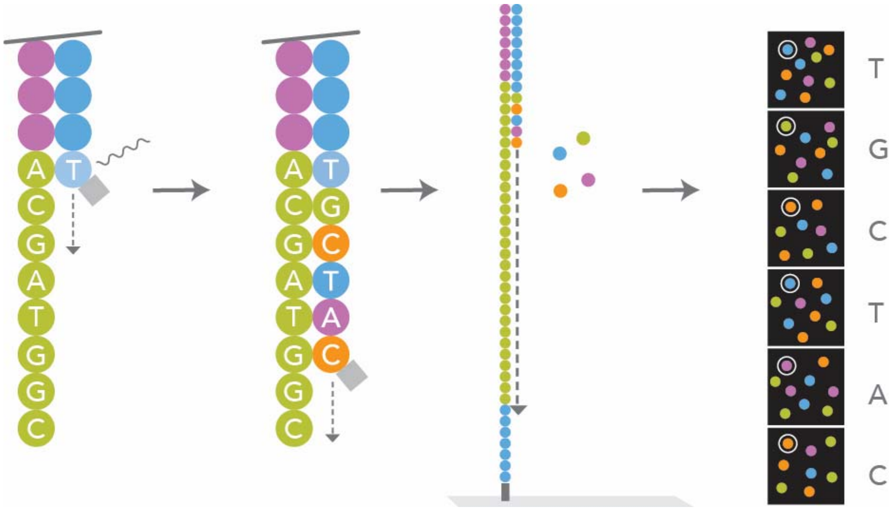


Figure 1.2 Illumina Sequencing by Synthesis technology. Courtesy of Illumina, Inc.

Illumina/Solexa develops sequencing by synthesis technology, which also uses fluorescence to detect when a base is added to a nucleotide sequence by a polymerase (Barski et al., 2007). Illumina technology uses a flow cell that allows DNA fragments with an adapter sequence, which also acts as a primer on the complementary strand, to attach randomly to its surface. The DNA fragments are exposed to all four dNTPs at once, with each dNTP connected to a differently coloured fluorophore, so that a new base can be added to the complementary strand by a polymerase. These dNTPs have a protecting group at their 3'-O-position so that further synthesis of the new strand is terminated (Turcatti et al., 2008). After a single base is added to a new strand, the colour of fluorescence corresponding to the newly added base is detected, and a wash cleaves the fluorophore and terminator sequence from the growing strand so another nucleotide can be added. The desktop Illumina machine, called MiSeq, produces 10-120Gb per run with read lengths from 50-300bp, while the most powerful machine, called the HiSeq X Ten, produces 900-1800Gb per run with read lengths from 50-150bp (Vincent et al., 2016). Table 1.1 shows read lengths and throughput produced by most of the sequencing technologies discussed in this thesis. A more extensive table of sequencing platforms was included in a review to mark the 10 year anniversary of NGS (Goodwin et al., 2016).

Table 1.1 Read length and throughput of various sequencing technologies

Sequencing Platform	Read Length	Throughput
454 GS FLX+	700 bp (mode)	700 Mb
Illumina MiSeq	50-300 bp	10-120 Gb
Illumina HiSeq X	50-150 bp	900-1800 Gb
Pacific Biosciences RS II	10 Kb (mode)	750 – 1250 Mb
Oxford Nanopore MinION	5.5 Kb (median)	Up to 1.5 Gb

1.1.4. Difficulty of closing a genome

Genomic reads are typically used to assemble longer genomic sequences. There are two main strategies for this: reference-based assembly, where reads are aligned to a pre-existing reference sequence, and *de novo* assembly, where reads are assembled without a reference sequence. A reference sequence is any nucleotide or protein sequence being used as a standard against which other sequences are being compared, and a reference genome is a genome being used as a standard against which other genomes are sequences are being compared. The basic strategy of *de novo* assemblers is to find reads with portions that align to each other, and to connect those reads together. The contiguous sequences that are produced from sequence assembly are referred to as contigs. Compared to eukaryote genomes, *de novo* assembly is used more often for bacterial genomes because the choice of reference sequence is less straightforward

(Pightling et al., 2014), and genome rearrangements or genomic regions that are present in the genome being assembled but absent from the reference genome are often of particular interest to researchers but would not be detected using a reference-based assembly (Hernandez et al., 2008). A bacterial genome sequence is considered complete or finished if the sequence of each chromosome is of high quality (with less than 1 error per 100 kb) and contained in a single contig, whereas in a draft genome sequence, which may be of high or low quality, the chromosome is contained within multiple contigs (Chain et al., 2009). While there is no agreed upon criteria in the international microbiology community for the definition of high quality draft genome, the Brinkman lab and others consider a draft genome to be of high quality if the overall coverage represents at least 90% of the genome. Automated sequence assemblers typically do not produce a complete bacterial genome from Illumina short read sequencing data.

A major factor that prevents closing of genomes is the presence of repetitive regions. Repetitive regions contain sequences that occur more than once in a genome. This definition applies to both short repetitive elements and longer sequences with multiple copies in a genome (Treangen & Salzberg, 2012). Unless a sequence read containing a repetitive sequence traverses its entire length, the alignment of the read to the genome is ambiguous (Schatz et al., 2010). A strategy to overcome this is paired end sequencing, where both ends of a DNA fragment are sequenced, and the approximate distance between the two ends can easily be estimated based on the fragment size (Roach et al., 1995). While paired reads do help overcome difficulties of sequencing repetitive regions, there are still limitations (Phillippy et al., 2008). Sequence assembly algorithms can handle reads with ambiguous alignments by ignoring them, randomly assigning them, or reporting multiple alignments (Treangen & Salzberg, 2012). This can create gaps between contigs, or misassemblies. Misassemblies are large-scale assembly errors, typically resulting from ambiguous alignments. Misassemblies can be rearrangements or inversions of genomic regions, or the collapse or expansion of repetitive regions (Phillippy et al., 2008). Strategies for the detection of misassemblies are discussed in section 2.1.

As well as the computational problem of repetitive regions, limitations of the sequencing process also prevent genome closing. The whole-genome amplification stage of sequence library preparation does not amplify the entire genome uniformly. Both tandem and inverted tandem repeats affect read coverage, and whether affected regions are over- or underrepresented depends on the amplification method used (Tsai et al., 2014). Both high and low GC regions are underrepresented, and while there is empirical evidence suggesting that PCR is the most important cause for this (Benjamini & Speed, 2012), some bias in coverage due to GC content is still present in unamplified sequencing libraries (Tsai et al., 2014). Sequencing errors also have a GC content bias: GC-rich regions, in particular regions with GGC motifs or G homopolymers, are particularly prone to sequencing errors in Illumina sequencers (Nakamura et al., 2011). Regions with inverted repeats are also more prone to errors (Nakamura et al., 2011).

1.1.5. Long read sequencing technologies

Sequence technologies that produce longer reads, and whose error profiles are unbiased, are more capable of closing genomes than short read sequencing technologies. There are two long read sequencing technologies that are currently being used in the genomics community: "Single Molecule, Real Time" (SMRT) technology from Pacific Biosciences (Flusberg et al., 2010) and nanopore technology from Oxford Nanopore (Clarke et al., 2009). SMRT technology, like sequencing by synthesis technology, measures the fluorescence of dNTPs labelled with four distinct colours as they are added to a growing complementary strand (in the case of SMRT sequencing, the fluorophore is linked to the terminal phosphate of the dNTP). However, unlike sequencing by synthesis technology, SMRT technology achieves single molecule resolution by binding the polymerase to the bottom of a nanophotonic structure called zero-mode waveguide, which allows for the detection of very small fluorescent emissions (Eid et al., 2009). Reads produced by SMRT technology have highly variable lengths, but half of the 750 Mb – 1.25 Gb of sequence produced by a SMRT cell is contained within read lengths of at least 20 kb (Pacific Biosciences, 2015).

Oxford Nanopore technologies sequence DNA using a nanometer scale pore that connects a salt solution with a voltage gradient. As negatively charged, single stranded DNA passes through the pore, the change in electrical current can be used to determine the specific nucleotide sequence (Laszlo et al., 2014). The origin of nanopore sequencing was several years ago (Kasianowicz et al., 1996), but relatively recent improvements including the use of Msp, an engineered porin protein (Derrington et al., 2010), the use of phi29 DNA polymerase to control translocation across the pore (Cherf et al., 2012), and improved algorithms for interpreting changes in current (Laszlo et al., 2014) have led to its commercialization. The MinION sequencer from Oxford Nanopore first became available to some researchers in 2014, and was released commercially in May 2015 (Oxford Nanopore, n.d.). Nanopore sequencing theoretically has no limit on read length, and reads of greater than 230 Kb have been sequenced with MinION (Ip et al., 2015). The median read length of the MinION according to the MinION analysis and reference consortium is 5.5 Kb, and the yield of a single run can be up to 1.5 Gb (Ip et al., 2015). Oxford Nanopore sequencers, and analyses of their output, are rapidly improving (Loman & Watson, 2015), so the user base of this technology is likely to grow.

Initially, long read sequencing was used in combination with short read sequencing technologies to help close genomes (Bashir et al., 2012; Ribeiro et al., 2012). Both long read sequencing technologies have since been used to produce complete bacterial genomes in a single run, with no need for additional sequencing. To overcome the high error rate of SMRT sequencing, the hierarchical genome assembly process (HGAP) incorporates a preassembly step where shorter reads are aligned to longer reads and used for error correction (Chin et al., 2013). This creates long, high quality sequences to be assembled. The HGAP preassembly step produced 17 232 reads with a mean length of 5 777 bp and a mean accuracy of 99.9% using reads from 8 cells (a single run) for an *E. coli* K12 MG1655 genome (Chin et al., 2013). A similar method was used to produce a complete version of the same genome using only MinION reads (Loman et al., 2015). While the ability to close a genome with a single sequencing technology has the potential to eventually eliminate the problem of draft genomes, the cost of these technologies and the investment that many institutions have already made into sequencing by synthesis technologies (Allard et al., 2016) mean that the problem of draft genomes is not yet eliminated.

1.2. Mobile genetic elements and genomic islands

1.2.1. Horizontal Gene Transfer

Horizontal gene transfer (HGT) is the transfer of genes between organisms by a mechanism other than direct inheritance. HGT is theorized to have occurred at very high frequencies in the early stages of life, allowing for new genes to spread easily through a population (Woese, 1998). Rates of HGT are still high in bacteria and archaea, and occur more frequently than single nucleotide substitutions in at least some bacteria (Hao & Golding, 2006). Ancient and modern HGT both complicate the production of phylogenetic trees. In particular, HGT has been problematic for the production of a single, universal tree of life (Doolittle & Bapteste, 2007). HGT in bacteria and archaea are also the basis of a major criticism of the species concept because different parts of an individual's genome can have different lineages (Ereshefsky, 2010). HGT remains an important mechanism for the acquisition of adaptive traits (Dobrindt et al., 2004; Sui et al., 2009). The importance of genomic regions acquired via HGT will be discussed in more detail in section 1.2.6. HGT can involve mobile genetic elements (MGEs), which encode mechanisms for their own transfer. There are three main mechanisms of HGT: transformation, conjugation, and transduction. These mechanisms will be briefly described in this section.

Transformation is the uptake of free DNA from the environment (Griffith, 1928). Competent bacteria, bacteria which are capable of undergoing transformation, maintain several genes that regulate competence in their genome. The mechanism DNA uptake is similar across most competent bacteria. Competent gram negative bacteria have an additional secretin channel which transports double stranded DNA (dsDNA) across the outer membrane, which is not needed in gram positive bacteria. When dsDNA reaches the inner membrane of gram negative bacteria, or the membrane of gram positive bacteria, nuclease or strand separating proteins separate the DNA into single strands. One strand passes through a transmembrane pore, and the other strand is degraded (Johnston et al., 2014). There are multiple theories regarding the reason bacteria maintain competence genes despite the risk of harmful mutations: it provides nutrition, it creates the potential for fast adaptation to the environment, or it can repair damaged DNA through homologous recombination (Finkel & Kolter, 2001; Redfield, 1988).

Conjugation is the transfer of DNA between two cells via direct interaction between a donor and a recipient (Lederberg & Tatum, 1946). Like transformation, conjugation requires many genes that are highly conserved (Alvarez-Martinez & Christie, 2009). DNA transfer and replication proteins, which include a relaxase and accessory factors, bind to a specific origin of transfer sequence, separate the two DNA strands, and facilitate the transfer of single-stranded DNA through a type IV secretion system (Alvarez-Martinez & Christie, 2009; Wozniak & Waldor, 2010). The transferred DNA may or may not encode the machinery required for its own transfer. Plasmids and elements which integrate into the genome can both be transferred by conjugation, and the latter is discussed in more detail in later sections.

Transduction is the transfer of DNA to a recipient cell through a bacteriophage (commonly called phage). When a phage enters the lysogenic stage, it integrates its own genome into the genome of the infected cell (Freifelder & Meselson, 1970), so the infected cell replicates the phage genome along with its own genome every time it divides. This usually involves integrating into the bacterial chromosome, but in some cases the phage genomes become circular or linear plasmids in the infected cell (Casjens, 2003). A phage chromosome that has integrated into a bacterial genome is called a prophage. Prophages are discussed in more detail in the next section.

1.2.2. Prophage

Phages are the most abundant, and possibly the most diverse, life forms on Earth (Suttle, 2005). Phages can be divided into two major groups: temperate phages, which are capable entering a lysogenic stage, and virulent phages, which are not (Lwoff, 1953). The former are more relevant to this work as they are the source of prophages, but virulent phages can also be an indirect source for HGT through recombination with a temperate phage (Fortier & Sekulovic, 2013). Most phage are specialized to infect a small subset of bacterial species, but some have a broader host range (Flores et al., 2011). Phage can acquire bacterial DNA from a host, and transfer that DNA to a new host. Specialized transduction can occur when a prophage is excised imperfectly such that flanking bacterial DNA is excised as well, whereas generalized transduction can occur if random bacterial

DNA from any part of the genome is accidentally packaged into a transducing phage (Canchaya et al., 2003).

Some prophages are inserted into random locations in the host chromosome, while others are inserted preferentially at certain sites via site-specific recombination (Campbell, 1992). These insertion sites may be within tRNA genes or other repetitive sequences. Phage genomes often contain tRNA genes that were probably acquired from one of their previous hosts (Bailly-Bechet et al., 2007). They are the only translation-associated genes found in phage genomes and despite their small genomes, some phages carry more than 20 tRNA genes. Maintaining tRNA genes might be beneficial for phage genomes by compensating for differences in codon usage bias between the phage and its host.

Phages are ubiquitous in the environment, and are estimated to infect 10^{23} bacteria per second in oceans alone (Suttle, 2007). They are also estimated to cause about half of all bacterial deaths in the ocean (Fuhrman & Noble, 1995). Phages are a major source of HGT; it is estimated that 10^{25} to 10^{28} bp of DNA is transferred in oceans by phages per year (Rohwer & Edwards, 2002). Prophages are very prevalent in bacterial genomes. Bacterial genomes may consist of up to 20% phage genes, and the majority of bacterial genomes contain at least one prophage sequence (Casjens, 2003; Paul, 2008).

1.2.3. Integrons

An integron is a genetic element capable of integrating DNA fragments, but does not encode a mechanism for genetic transfer (Stokes & Hall, 1989). Some integrons are part of a bacterial chromosome, while others are encoded on plasmids (Escudero et al., 2015). Integrons contain two main components: cassettes, which are variable, and a platform, which tends is more highly conserved. The platform contains elements required for integration and expression of cassettes: a tyrosine recombinase gene and its promoter, a recombination site for cassette integration, and a promoter to activate expression of the cassette. The tyrosine recombinase is highly conserved, but the recombination site is a more variable inverted repeat sequence (Bouvier et al., 2005). An integron can contain more than one cassette; in some cases, an integron can accumulate hundreds of cassettes, such as the *Vibro cholerae* superintegron which can take up 3% of the whole genome (Escudero et al., 2015; Mazel et al., 1998). Integrons are present in about 17% of sequenced bacterial genomes from a variety of taxonomic groups, but they are particularly prevalent in freshwater proteobacteria and marine ɣ-proteobacteria (Cambray Guillaume et al., 2010).

1.2.4. Transposons and IS elements

Transposons are genetic elements that are able move to different locations in a genome. Transposons were first identified in maize (McClintock, 1941), and they can be found in genomes from all three domains of life (Langille et al., 2008b). They are quite diverse, and can range in size from hundreds to more than 65 000 bps. Transposons may either be autonomous, meaning they encode a mechanism to transpose themselves, or non-autonomous, meaning that they must rely on other transposition machinery. Transposases are proteins which catalyze transposition of DNA. There are several unrelated families of transposases, and different transposases can have different integration specificities. Many prefer to integrate into the 3' end of tRNA genes or other conserved genes, but others have low integration specificity (Bellanger et al., 2014). Most transposons have terminal inverted repeat (TIR) sequences at both ends which can act as transposase binding or cleavage sites (Langille et al., 2008b; Mahillon & Chandler, 1998).

Conjugative transposons encode machinery for their transfer by conjugation as well as encoding transposition machinery (Roberts et al., 2008). Besides conjugation and transposition machinery, the gene content of conjugative transposons is highly variable, even within transposon families which share closely related conjugation and transposition machinery (Bellanger et al., 2014). Most studied conjugative transposons are able to transfer themselves at least between closely related genera, but some conjugative transposons, such as the well studied Tn916 which confers tetracycline resistance, are able to transfer between diverse bacteria.

Insertion sequence (IS) elements are small MGEs, 0.7 to 2.5 kb, which typically only contain genes required for their own mobility and TIR sequences (Adhya & Shapiro, 1969; Darmon & Leach, 2014; Shapiro & Adhya, 1969; Shapiro, 1969). Like transposons in general, IS elements have diverse transposases, and have varying degrees of insertion site specificity (Mahillon & Chandler, 1998).

1.2.5. Genomic Islands

Genomic islands (GIs) are segments of bacterial or archaeal chromosomes, consisting of several genes, that have probable horizontal origins (Langille et al., 2010). They are typically defined as being at least 8 kb in length. This length restriction serves as a practical cut off for use in GI prediction, and it excludes shorter MGEs such as IS elements. The term also excludes MGEs that do not integrate into the chromosome such as plasmids. Figure 1.3 is a diagram representing which MGEs are considered GIs, and which are not. GIs have previously been referred to as pathogenicity islands due to the ability of some GIs to induce a pathogenic phenotype in bacteria which carry them (Hacker et al., 1990). Several names were thereafter used to describe genetic elements that were similar to pathogenicity islands, but encoded different functions. These names include metabolic island, antibiotic-resistance island, and symbiosis island (Dobrindt et al., 2004). GI is an umbrella term used to describe these genomic regions regardless of the phenotypic traits they confer.

Some definitions of GIs include a requirement that they are absent from the genomes of closely related strains (Darmon & Leach, 2014). This is a logical requirement because GIs are horizontally acquired regions, and this requirement has been used to support the validity of comparative genomics approaches to identify GIs (Karaolis et al., 1998). However, this definition should be interpreted with caution because comparative genomics approaches may identify different GIs depending on the strains used for comparison (Langille et al., 2008a). GIs may have been acquired very recently, or they may be ancient. Comparative genomics approaches for GI prediction, and the difficulty of predicting ancient GIs, will be discussed in more detail in section 1.3.

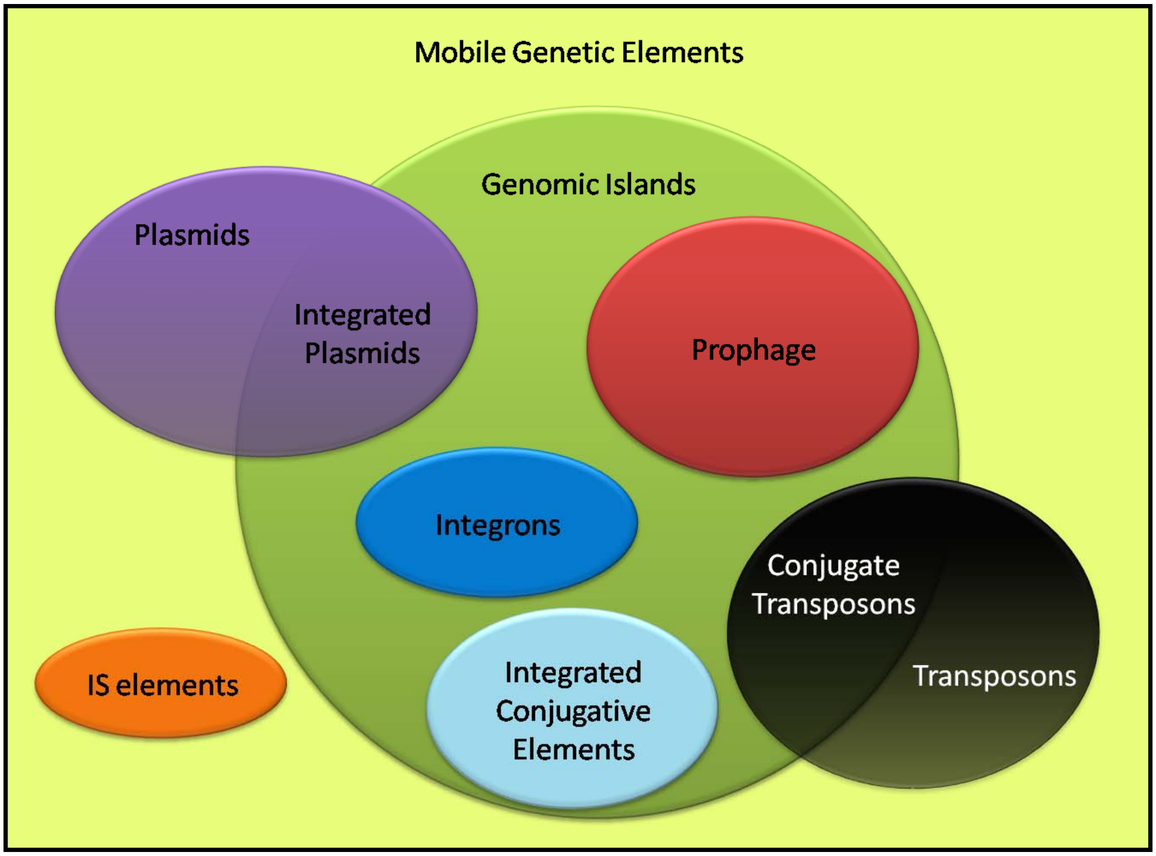


Figure 1.3 Different kinds of MGEs, and which MGEs are considered GIs. Figure is from Langille, 2009 (Langille, 2009) **and is used with permission of the author.**

1.2.6. Features of GIs and interest in GIs

GIs have a different gene composition than non-GI regions. GIs disproportionately contain novel genes, or genes with no known homologs (Hsiao et al., 2005). This may be because GIs originate from a very large gene pool which includes the phage gene pool. As mentioned earlier, phages are possibly the most diverse life forms on Earth, yet they are not as well studied as other life forms. Unsurprisingly, mobility genes such as those encoding transposases are also disproportionately found in GIs. GIs are also associated with carrying a variety of genes for environmental adaptations such as those which enable the degradation of xenobiotic compounds and synthesis of polyketides (Dobrindt et al., 2004). Acquisition of antimicrobial resistance (AMR) is often associated with either GIs or plasmids, and integrons are the main mechanism for acquisition of AMR by Gram-negative enterobacteria. Virulence factors (VFs), which are genes or other characteristics of bacteria associated with the ability to cause disease, are also associated with GIs. VFs are disproportionately found in GIs across many bacteria, especially those that are capable of inhabiting multiple environmental niches (Sui et al., 2009). Type II and type IV secretion system components, toxins, and adherence factors in particular are overrepresented in GIs.

There are several well documented examples of GIs, and prophages in particular, contributing to virulence (Fortier & Sekulovic, 2013). The namesake for pathogenicity islands in uropathogenic *E. coli*, which contained a hemolysin determinant and genes encoding a fimbriae, was able to convert a non-pathogenic strain to a pathogenic one (Hacker et al., 1990). Many well characterized toxins are encoded in prophages, including botulinum toxin in *Clostridium botulinum,* (Eklund et al., 1971), diphtheria toxin in *Corynebacterium diphtheriae* (Freeman, 1951), and Shiga toxin in *Shigella dysenteria* and some strains of *E. coli* (O'Brien et al., 1984). There are documented examples where antibiotic use has caused dormant phages to enter lytic stage, causing increased expression of phage-encoded genes, and increased virulence (Zhang et al., 2000). Prophage genes are highly expressed in *Pseudomonas aeruginosa* biofilms, and biofilms can allow for persistent infections in the lungs of cystic fibrosis patients (Whiteley et al., 2001).

Analysis of GIs, along with SNP-based analysis, can also help resolve transmission patterns. A recent example of the usefulness of GI analysis was a *Listeria monocytogenes* outbreak in Sydney, Australia where epidemiological data had traced the source to a hospital food supplier (Wang et al., 2015). Traditional typing methods and SNP-based analyses were insufficient to link clinical outbreak isolates with isolates from the food supplier, but when SNP data was combined with GI data the researchers were able to confirm the suspected epidemiological link. This case highlighted the importance of GI analysis for genomic epidemiology.

Within-patient recombination is also of particular medical interest, although this area of research is still in its infancy. There has been disagreement about the rates of HGT between *P. aeruginosa* cells within cystic fibrosis lungs with persistent *P. aeruginosa* infections, and to what degree HGT contributes to within-host diversity (Williams et al., 2016). Rates of HGT within the gut microbiome are estimated to be high, driven by the dense and diverse environment (Smillie et al., 2011), but recombination in the gut microbiome is harder to study because of the complexity of the microbial community.

In conclusion, GIs are of particular interest to researchers due to both their unique implications to evolution and their disproportionate contributions to disease.

1.3. Genomic island prediction

1.3.1. Sequence composition based methods

GIs can have certain compositional signatures that differ from their host genome, and these differences can be used to computationally detect GIs within a microbial genome sequence. These compositional differences include GC content, codon usage, and oligonucleotide frequency, which are guided by selective constraints in the host genome (Burget et al., 1992; Sueoka, 1992). GIs are not under the same selective constraints until after entering the host genome, and MGEs may have different selective pressures prior to entering a host genome which lead to distinct compositional signatures (Rocha & Danchin, 2002). As discussed in section 1.2, gene composition also differs in GIs, with certain gene types such as transposase and tRNA genes being overrepresented in these regions and can also be used in combination with sequence compositional bias for GI prediction. Two advantages of sequence composition based methods for GI prediction are that they do not require wet lab experiments, and unlike comparative genomics approaches (which are discussed in section 1.3.2) they do not require similar comparison genomes. A major limitation of these methods is that not all GIs have strong compositional signatures; many GIs are acquired from closely related organisms that have similar compositional signatures, and ancient GIs may have evolved with their host and lost their previous differences (Langille et al., 2010).

Some sequence composition based methods use hidden Markov models (HMMs). HMMs are statistical models that can be applied to a wide variety of observable features with underlying states (Blunsom, 2004). In molecular biology or GI prediction, these observable features can include gene composition, or raw amino acid or nucleotide sequences. With supervised training, HMMs are built from examples where observable features (such as amino acid sequence) and underlying states (such as the protein product) are both known. HMMs built from multiple sequence alignments are called profile HMMs (Eddy, 2003), which can be represented as position-specific scoring matrices (PSSMs). There are two BLAST tools designed specifically for use with HMMs: PSI-BLAST performs searches using an HMM profile against a protein database, and RPS-BLAST performs searches using a protein sequence against a PSSM database (Altschul et al., 1997). HMMER is a tool for building and rapidly searching HMM profiles (Finn et al., 2011), and is incorporated into multiple sequence composition based GI prediction methods.

There are many sequence composition based GI prediction methods, both with and without HMMs. SIGI-HMM (Waack et al., 2006) uses a HMM to detect GIs based on codon usage bias. PIPs (Abreu et al., 2012) and its extension, GIPSy (Soares et al., 2015), incorporate SIGI-HMM into their GI prediction pipelines, but they also detect atypical GC content and gene content associated with GIs. Another approach calculates Markovian Jensen-Shannon divergence of genomic regions, which generally represents nucleotide frequencies, for regions of decreasing size with a recursive algorithm until it detects atypical regions (Arvey et al., 2009). Alien Hunter (Vernikos & Parkhill, 2006) uses oligonucleotide frequencies to detect GIs, and uses an HMM to determine optimal boundaries of predictions. PredictBias (Pundhir et al., 2008) predicts sequences with either dinucleotide or GC content bias as putative GIs, and performs a RPSBLAST search of VF profiles to determine whether a GI is a pathogenicity island. IslandPath-DIMOB (Hsiao et al., 2005) detects GIs using dinucleotide biases, and requires the presence of a mobility gene, either within user-provided annotations or detected using HMMER3, to consider a region with dinucleotide bias as a GI. Centroid (Rajan et al., 2007) and INDeGenIUS (Shrivastava Sakshi, Ch V Siva Kumar Reddy, 2010) are other methods which detect differences in oligonucleotide frequencies, but do not use HMMs. MGSIP (de Brito et al., 2016) uses a mean shift clustering algorithm to detect GIs based GC content bias. PAI-IDA (Tu & Ding, 2003) detects GIs based on GC content, codon usage bias, and dinucleotide frequency. Z-island explorer is a web-based GI prediction tool which uses GC content variation to detect GIs (Wei et al., 2016).

Since there are many features and methods that can be used for GI detection, it is important to consider which methods are most appropriate for different purposes. A previous analysis of sequence composition based GI prediction methods used results from a comparative genomics method for GI prediction (IslandPick, described in the next section) to evaluate the performance of many tools using 117 microbial genomes (Langille et al., 2008a). This evaluation found that AlienHunter had the best sensitivity, SIGI-HMM was the most precise, and IslandPath-DIMOB and SIGI-HMM had the highest overall accuracy. It was therefore recommended that researchers who are willing to sort through false positives in order to have a high recall may prefer AlienHunter, while researchers who would prefer to have a small but accurate list of GIs with the risk of false negatives may prefer SIGI-HMM or IslandPath-DIMOB. By combining the latter two methods, there was an increase in recall with minimal decrease in sensitivity.

1.3.2. Comparative genomics methods

The first genetic element to be termed a pathogenicity island was discovered by comparing related strains of *E. coli*, where the genomes of some isolates contained this element while others did not (Hacker et al., 1990). This is an early example of a comparative genomics-based approach to GI detection, although in this case the genome sequences were not available and the comparison method was not computational. The basic approach of comparing an isolate of interest with related isolates, or a combination of this approach with sequence composition based methods, is still the preferred method for GI prediction when appropriate isolates are available for comparison (Langille et al., 2010). Comparative genomics approaches identify GIs as regions of a query genome that are absent from related isolates, so results are highly dependent on the genomes used for comparison. If only very closely related genomes are used for comparison, only very recently acquired GIs will be detected. If very distantly related genomes are used, the risk of false positives due to chromosomal rearrangements increases. Consideration should therefore be given to choosing appropriate genomes for comparison.

There are multiple computational methods that use a comparative genomics approach to predict GIs. DarkHorse (Podell & Gaasterland, 2007) performs a BLAST search of all predicted proteins in a genome against the NCBI non-redundant protein database, uses the results to generate lineage probability index scores, and reports predicted proteins with highly ranked scores as having possible horizontal origins. The MobilomeFINDER (Ou et al., 2007) comparative analysis tool, which requires that users manually select comparison genomes, detects regions that are unique to the query genome using Mauve, a whole genome multiple sequence aligner (Darling et al., 2004). These unique regions are then filtered by requiring that they are flanked by tRNA gene segments, meaning that MobilomeFINDER only reports GIs which have inserted into tRNA genes.

IslandPick (Langille et al., 2008a) also uses Mauve in its comparative genomics approach to GI prediction. IslandPick performs pairwise alignments against comparison genomes, extracts unique regions of greater than 8 kbp which are present in the query but absent from all comparison genomes, and performs a BLAST search of these regions against the query to filter out genome duplications. A major benefit of IslandPick is that it includes a method for automated selection of comparison genomes. CVTree (Xu & Hao, 2009) is used to calculate genetic distances between the query genome and a set of possible comparison genomes, and uses a specific set of selection criteria to select, when possible, three to six genomes for comparison. The selection criteria include minimum and maximum distances from the query genome, as well as single close genome and single distant genome cut-offs that ensure not all selected genomes are very close to or very distant from the query genome.

1.3.3. Databases and other computational resources

There are several databases for GIs, or certain subsets of GIs, which include search tools to detect GIs in a user-provided DNA sequence. The Islander database (Hudson et al., 2015) specifically contains predicted GIs that have inserted into a tRNA gene and encode an integrase, and the program used to detect these predicted GIs is available for download on the database website. PAIDB (Yoon et al., 2014) is a database of pathogenicity and AMR islands which includes a tool for BLAT and BLASTx-based GI prediction. ICEberg (Bi et al., 2011) is a database specifically for ICEs, and has an option BLAST and HMMER3 searches on its web interface. MOSAIC (Chiapello et al., 2008) is a comparative genomics database containing information about conserved and variable genomic regions developed using pre-computed whole genome alignments. HGT-DB (Garcia-Vallve et al., 2003) is a database of sequence composition information (GC content and codon usage) for genes in bacterial and archaeal genomes, including whether the sequence composition of the gene deviates from the rest of the genome.

1.3.4. IslandViewer

IslandViewer is a web-based GI prediction tool that incorporates three of the most accurate GI prediction methods: SIGI-HMM, IslandPath-DIMOB, and IslandPick (Hsiao et al., 2003; Langille & Brinkman, 2009; Langille et al., 2008a; Waack et al., 2006). SIGI-HMM and IslandPath-DIMOB were identified as the most accurate sequence composition based prediction methods in a previous evaluation (see section 1.3.1), and IslandPick uses a comparative genomics approach for GI prediction (see section 1.3.2). IslandViewer contains pre-computed results from these methods for all complete bacterial and archaeal genomes available on NCBI, managed by the MicrobeDB database (Langille et al., 2012). Users can also upload annotated genomes to IslandViewer for custom GI predictions. IslandViewer results can be viewed using a circular genome map, and results can be downloaded in multiple file formats. After the first IslandViewer update, curated VF and AMR genes could also be viewed and downloaded for a subset of pre-computed genomes, along with a previous analysis of pathogen-associated genes. Pathogen-associated genes are only ever detected in pathogens and never in non-pathogens sequenced to date, and may contribute to virulence (Dhillon et al., 2013).

Developed by the Brinkman laboratory, IslandViewer is freely available without a subscription and widely used. As of July 2016, the IslandViewer papers have 277, 41, and 32 references according to Google Scholar (Dhillon et al., 2013, 2015; Langille & Brinkman, 2009), and hundreds of genomes are uploaded for custom GI prediction every month. Since its creation, IslandViewer has had multiple updates to increase its functionality and to provide flexible and interactive visualizations of results (Dhillon et al., 2013, 2015), and more updates are underway. In particular, an improved version of IslandPath-DIMOB with increased overall accuracy will soon be implemented into IslandViewer. Chapter 3 will discuss several recent improvements to IslandViewer in more detail, with particular emphasis on a new draft genome analysis pipeline.

1.4. Prediction of other genomic regions of interest

1.4.1. Antimicrobial Resistance genes

The standard methods used for AMR detection are still based on culturing and measurement of minimum inhibitory concentration (Jorgensen et al., 2009), but several computational methods have recently been developed to detect AMR determinants from a bacterial genome sequence using a database of known AMR genes. ResFinder (Zankari et al., 2012) is the name of both a database and a corresponding AMR detection tool. The detection tool essentially performs a BLAST search of an assembled genome against the database. ARG-ANNOT (Gupta et al., 2014) is also a database and AMR detection tool. The ARG-ANNOT detection tool uses a BLAST search against a database of AMR proteins, but also searches for single nucleotide variants (SNVs) which confer AMR. SRST2 (Inouye et al., 2014) and SEAR (Rowe et al., 2015) both contain tools for AMR detection from unassembled sequence reads, and they both use the ARG-ANNOT database. The Comprehensive Antibiotic Resistance Database (CARD, McArthur et al., 2013) is, as its name suggests, a comprehensive AMR database which is continually updated. As of July 2016, CARD contains 2433 AMR gene sequences and 963 SNVs (http://card.mcmaster.ca). The Resistance Gene Identifier (RGI) tool performs a BLASTp search of CARD with precise filtering criterion, and identifies SNVs which confer resistance. Resfams (Gibson et al., 2015) detects AMR genes by searching translated open reading frames (ORFs) against hidden Markov models (HMMs) of AMR protein families using hmmscan (Finn et al., 2011). The protein family models were developed in part from CARD. Large, well curated databases increase the efficacy of computational AMR prediction methods which depend on them, and in turn, computational AMR prediction can expand the number of known AMR genes (Gibson et al., 2015).

1.4.2. Virulence Factors

Computational VF prediction also relies largely on databases of known VFs. The Virulence Factor Database (VFDB, (Chen et al., 2012)) consists of two main data sets: one that consists solely of experimentally verified VFs from 24 medically important bacterial genera, and another that also includes VF homologs using reciprocal BLAST searches. The VFDB web interface allows users to perform a BLAST search of either data set. Victors virulence factors (http://www.phidias.us/victors/) is a database that consists solely of VFs that have been experimentally observed, and currently contains VFs from 194 pathogens. mVirDB (Zhou et al., 2007) is a database that combines toxin, VF, and AMR genes from various sources. The PATRIC database (Wattam et al., 2014) contains various gene types associated with pathogenic organisms, including VFs. The web interfaces for each of these databases include a BLAST search tool.

1.4.3. Functional categories

Proteins can be grouped into functional categories or families, and there are multiple databases that can be used to assign a protein to a functional category. KEGG, a set of databases containing a range of biological information including pathways, contains a GENES database for cross-species annotations of genes and proteins, and a BRITE functional hierarchy of these genes (Moriya et al., 2007). The KAAS genome annotation tool can assign protein coding sequences (CDSes) to KEGG GENE entries and infer information about functional pathways. Pfam (Bateman et al., 2004) is a database of protein domain families. Each family has been manually curated, but additions to the family are found automatically through HMMER3 searches during updates (Finn et al., 2011). Pfam profile HMMs are built from curated seed alignments, and can be used to detect protein families from protein sequences. Sets of families that are homologous but too highly diverged to be classified as a single family are called clans (Finn et al., 2006). A Cluster of Orthologous Groups of proteins (COG) is a group of proteins from at least three distantly related species that are predicted to belong to an orthologous protein family (Tatusov et al., 1997). Position-specific scoring matrices have been created for each COG category, which can be used to assign protein sequences to COG categories (Marchler-Bauer et al., 2009). COGs are classified into at least one of 25 superfamilies which broadly describe their function (or lack of known function) (Tatusov et al., 2000).

1.5. Goals of present research

The Integrated Rapid Infectious Disease Analysis project (IRIDA, http://www.irida.ca) is a collaboration between the Public Health Agency of Canada, the

British Columbia Centre for Disease Control, and the Brinkman Laboratory at Simon Fraser University to develop a bioinformatics platform for infectious disease genomic epidemiology. For GI prediction in outbreak isolate genomes, IslandViewer is being incorporated into IRIDA. The current version of IRIDA assembles genomes using Illumina read data and tends to produce draft genomes, so in order for IslandViewer to be integrated into IRIDA, it is essential for IslandViewer to be able to accept draft genomes as input. As analyses using draft bacterial genomes become increasingly common, including in clinical and epidemiological settings, it is also important to understand the characteristics of draft genomes compared to complete genomes. At the onset of my project, there had been relatively little research which characterized the utility of draft bacterial genomes for functional analysis. There was also only one available tool for GI prediction which could accept a draft genome as input, and this tool became unavailable shortly after publication (Lee et al., 2013). In order to address these issues, the main goals of my project have been to perform an initial characterization of the importance of missing regions of draft genomes, with a particular focus on gene functional analysis and GI prediction, and to increase the functionality of IslandViewer software to allow for the analysis of draft genomes.

Chapter 2.

Characterization of missing regions in draft genomes

Note: I led the characterization of missing regions in draft genomes and performed the analysis. Bhavjinder Dhillon detected virulence factors in complete genomes, and Dr. Claire Bertelli provided genomic island predictions from one of the methods used: an updated version of IslandPath-DIMOB. For antisense transcription analysis, I developed the analysis pipeline, and Ogan Mancarci went on to do the analysis.

2.1. Background and rationale

2.1.1. Draft genome assessment background

As described in section 1.1.5, the vast majority of bacterial genome sequences are only completed to the draft stage. There have been multiple studies that have characterized draft genomes, and most of these characterizations have so far been in the form of evaluation of assemblies. Assembly evaluations focus on comparing traits such as number of contigs, contig lengths, N50 and variants of N50, and GC content. In a genome assembly, the N50 is the contig length such that 50% of the base pairs in the assembly are contained within contigs of that length or larger. In cases where a complete reference genome is available, they may also look at other factors such as the number of misassemblies. The first examples of these studies were Assemblathon and GAGE (Earl et al., 2011; Salzberg et al., 2012). Assemblathon was a contest where competing teams assembled simulated Ilumina HiSeq reads of a simulated genome evolved from human chromosome 13. GAGE is a tool for assembly assessment which requires a complete reference genome. GAGE was originally used to compare several assembly algorithms for their ability to assemble four genomes: *S. aureus*, *R. sphaeroides*, human, and bumble bee. A bacteria-specific version of GAGE, GAGE-B, was tested with eight bacterial genomes of varying size and GC content (Magoc et al., 2013). The first GAGE study found that ALLPATHS-LG performed well overall, and bacteria-specific study found that MaSuRCA and SPAdes produced the best assemblies, but in both studies the authors emphasized that each assembler had strengths and weaknesses, and that many assembly algorithms were still under development. Later assembly evaluations include QUAST (which was notably the first evaluation tool which did not require a complete reference genome), and iMetAMOS (Gurevich et al., 2013; Koren et al., 2014). Both of these tools found that SPAdes was the best assembler for the bacterial genomes which they assessed.

Another strategy for draft genome evaluation is the detection of misassemblies without a reference sequence. Misassembly detection algorithms align raw paired-end reads to *de novo* assemblies that were generated using those reads, and use the read alignments to generate a probability or likelihood score for every base in the assembly. These scores are then used to detect regions of the draft genome that are misassembled (Clark et al., 2013; Hunt et al., 2013; Rahman & Pachter, 2013). With these algorithms, regions to be resequenced can be identified, and a corrected N50 can be calculated without a reference sequence. These algorithms take into account issues surrounding repetitive genomic regions, provide an additional metric of genome quality for draft genomes without a reference, and are useful for improving the quality of the draft genome being assessed.

The draft genome assessments described above do not focus on gene content or characteristics of draft genomes that are directly pertinent to common uses of draft genomes. The extent of gene analysis in the above assessments is to calculate the total number of genes in the draft genome (Gurevich et al., 2013). More recently, there have been two studies that have assessed the quality of draft genomes based on the presence or absence of a set of marker genes. One study identified sets of universal single-copy orthologs for six major phylogenetic clades, and detected genes from this set that are duplicated, fragmented, or missing in a draft genome (Simão et al., 2015). Another study, which focused on microbial genomes, defined marker genes as any gene present in a single copy in at least 97% of genomes in a certain lineage (Parks et al., 2015), and used these marker genes to calculate the percent completeness for thousands of genomes.

To my knowledge, there has been only one previous study to assess gene functional categories in draft genomes (Klassen & Currie, 2012). This study focused on the fragmentation of open reading frames (ORFs) at contig boundaries. Gene fragmentation in draft genomes can lead to under-annotation if a gene fragment is not recognized or is mistaken for a truncated gene, but it can also lead to over-annotation in cases where a single gene is contained within two contigs, and the gene is annotated in each contig. Twenty-five *Streptomyces* were assessed for over- or under-representation of certain gene types based on Pfam, COG, and KEGG classifications. Based on these genomes, gene fragmentation led to over-annotation with KEGG and under-annotation with Pfam. Three COG superfamilies were identified as substantially enriched in fragmented ORFs, and this trend was driven mostly by 3 families: polyketide synthase modules and related proteins, non-ribosomal peptide synthetase modules and related proteins, and serine/threonine protein kinases. This study only addressed genes that were fragmented in draft genomes, and did not address genes that were missing from draft genomes.

2.1.2. Antisense transcription background

Cis-antisense transcription is the production of an RNA molecule copied from the sense strand of an ORF, such that the RNA molecule does not encode a functional protein. Antisense RNA molecules are usually small (relative to the length of the ORF), and are known to control gene expression in at least some cases (Georg & Hess, 2011). Early documented cases of antisense transcription disproportionately occurred in mobile elements (Wagner & Simons, 1994), and a personal communication indicated that there may be a bias in antisense transcription in genomic islands (GIs), but to my knowledge there has been no previous genome-wide assessment of the association of genomic islands with antisense transcription. Therefore, as well as a characterization of GIs and other features in draft genomes, this chapter also investigated whether there was any association of antisense transcription with GIs.

2.1.3. Rationale

While there has been extensive work on the evaluation of draft genomes, there is relatively little existing research that directly measures the limitations of draft genomes for comparing gene content. To my knowledge, there has been no assessment of the quality of draft genomes in terms of how useful they are for analyses such as antimicrobial resistance (AMR) gene and virulence factor (VF) prediction, gene function analysis, or GI analysis. This chapter presents research that begins to address this knowledge gap.

To characterize the importance of missing regions in two sets of draft genomes, draft bacterial genomes produced using Illumina sequencing by synthesis technology were compared with the subsequently completed genomic sequence from the same isolate. The main data set used for this analysis consists of thirty-six *Listeria monocytogenes* genomes sequenced by the Canadian National Microbiology Laboratory, where pairs of draft and complete genomes from the same isolates were used. Genomes of *Pseudomonas aeruginosa* reference panel isolates were also used as a secondary data set. *Listeria* and *Pseudomonas* have very different evolutionary lineages and genome characteristics (Glaser et al., 2001; Stover et al., 2000). *Pseudomonas* is Gram-negative, while *Listeria* is Gram-positive. *P. aeruginosa* genomes range from 5.5 to 7 Mbp and are GC rich, while *L. monocytogenes* genomes are typically about 3 Mbp and are AT rich.

L. monocytogenes and *P. aeruginosa* are both well studied pathogens, but the illnesses they cause and the mechanisms through which the cause illness are very different. *L. monocytogenes* is a foodborne pathogen, and it is the causative agent of listeriosis. Listeriosis can cause a variety of symptoms including gastroenteritis and septiciaemia, which is highly lethal (Hamon et al., 2006). *L. monocytogenes* is an intracellular pathogen, and it enters host cells by binding to receptors using internalin A or internalin B. Internalin A and internalin B bind proteins on the surface of a host cell, which causes host cytoskeletal rearrangement and the entry of *L. monocytogenes* into the host cell within a phagosome. After entering a host cell, *L. monocytogenes* releases itself from the phagasome by secreting phospholipases and listeriolysin O, a pore-forming toxin. After releasing itself from the phagosome, *L. monocytogenes* can use ActA to polymerize host actin, propel through the host cytoplasm, and pass into neighbouring host cells in a double membraned vacuole via membrane protrusion.

P. aeruginosa is an opportunistic pathogen which most commonly infects the lungs of cystic fibrosis patients and burn wounds (Gellatly & Hancock, 2013*).* As a species, *P. aeruginosa* is known for being metabolically diverse, for expressing a wide variety of VFs, and for causing persistent infections. VFs include flagella and type IV pili, which facilitate cell mobility and attachment to host epithelial cells, a type III secretion system, proteases, and lipopolysaccharide. When a local population is sufficiently high, *P. aeruginosa* can form a biofilm, which contributes to the persistence of infection. Each cell secretes autoinducer molecules, and when the local concentration of autoinducer molecules passes a threshold, *P. aeruginosa* cells in the area undergo changes in gene expression which results in biofilm formation.

Analysis of clusters of orthologous groups of genes (COGs), antimicrobial resistance genes, and virulence factors in regions present in the complete genome and missing from draft genomes was performed. The ability to detect GIs in draft genomes was assessed using IslandViewer. Together, these analyses show that there are limitations to bacterial draft genome analysis, with respect to disproportionately missing certain types of genes, however, valuable information of medical interest can still be obtained from some draft genome datasets.

2.2. General methods

2.2.1. Genome annotation and alignment of draft genome to complete genome

Draft and complete genomes were annotated using Prokka (Seemann, 2014), a widely used tool that combines several previously existing annotation tools in order to provide automated annotation of bacterial genome sequences. Both data sets used the default settings, but the *L. monocytogenes* data set was annotated with Prokka 1.7, while the *P. aeruginosa* data set was annotated using Prokka 1.11. A newer version of Prokka was used for the *Pseudomonas* data set in order to coordinate these analyses with other work being done on the same data set. Differences between these versions of Prokka are not expected to have a significant impact on the results of this study. Coding sequence (CDS) and tRNA gene annotation are the components of Prokka annotations that are used for this analysis. Prokka uses ARAGORN (Laslett & Canback, 2004) to predict tRNA genes. ARAGORN predicts tRNA genes by searching the genome for a small, conserved segment of the B-box promoter signal and searching for a tRNA structure around each initial hit. Prokka uses Prodigal (Hyatt et al., 2010) to predict CDSes, and uses a series of searches against increasingly broad databases to identify a putative gene annotation for each CDS.

Contigs from each draft genome were aligned to the corresponding complete genome from the same isolate, or complete genomes from closely related isolates in the case of most genomes in the *P. aeruginosa* data set, using Mauve Contig Mover (Rissman et al., 2009). An additional MegaBLAST step was performed where contigs that were not aligned with Mauve were aligned to the complete genome. Contigs that had 90% identity with a region of the complete genome over at least 90% of its length, with gaps of no more than 10% of its length were considered to correspond to that region of the complete genome, given that only one region of the complete genome met these requirements. The length limits were chosen empirically. A non-redundant list of coordinates covered by contigs was generated from the Mauve and MegaBLAST output using a custom script. Coordinates of genomic regions of interest were compared against this non-redundant list to determine whether these regions were present in the draft genome.

2.2.2. Identification of COGs and genes of interest

COG categories were assigned to CDSes as described previously (Klassen & Currie, 2012). COG motifs were retrieved from NCBI and used to create an RPSBLAST database. An RPSBLAST search with an expectation value cutoff of 0.00001 was performed on each CDS in each complete genome, and the top hit was assigned as the COG category corresponding to a given CDS.

VF and AMR gene homologs were predicted using the same methods that were used for genome annotation in IslandViewer 3 (Dhillon et al., 2015). For each complete genome in this study, Reciprocal BLAST searches were performed: all CDSes in the query genome underwent a BLASTp search against a database of all CDSes in a curated reference genome, and vice versa. Curated reference genomes were chosen such that the CVTree (Xu & Hao, 2009) distance between the curated and uncurated genomes was less than 0.3. Reciprocal best blast hits where the reference gene has a virulence factor annotation are then used to annotate virulence factors in the query genome. *L.*

monocytogenes EGD-e (NC_003210.1) was used as a reference genome for the *L. monocytogenes* data set, and *P. aeruginosa* PAO1 (NC_002516.2) and PA14 (NC_008463.1) were used as reference genomes for the *P. aeruginosa* data set. Antimicrobial resistance genes were annotated using the Resistance Gene Identifier (RGI) (McArthur et al., 2013), a tool linked to the CARD database which was described in section 1.4.1.

2.2.3. Genomic Island detection

GIs were predicted in both draft and complete genomes using IslandViewer 3 (Dhillon et al., 2015). For draft *L. monocytogenes* genomes, reference genomes used for contig arrangement were chosen by performing a BLAST search of the longest contig against all complete bacterial genomes in RefSeq as of September 18, 2014. Genomes within this data set were excluded from the BLAST search in order to simulate a data set where the complete genome from the same isolate is not available. For draft *P. aeruginosa* genomes, the complete RefSeq genome separated from a draft genome by the shortest total branch length based on a phylogenetic tree of *P. aeruginosa* genomes generated using parSNP (Treangen et al., 2014) was chosen as a reference genome. For this analysis, reference genomes were limited to RefSeq genomes that had been complete as of March 2015, as these genomes were readily available for use as reference genomes in IslandViewer.

As mentioned in chapter 1, a new version of IslandPath-DIMOB has been developed and will soon be incorporated into IslandViewer. In order to incorporate the most recent version of the IslandViewer prediction methods into this analysis, IslandPath-DIMOB results from IslandViewer were replaced with output from the most recent version of this method.

2.3. Draft *Listeria monocytogenes* genomes

2.3.1. The data set

The main data set for this analysis consists of 36 *L. monocytogenes* isolates where both the original shotgun sequencing reads, and the subsequently completed genome sequences from the same isolate, were provided by the Public Health Agency of Canada. Two of these genomes had been previously analyzed (Gilmour et al., 2010) and can be found on GenBank with the accession numbers NC_013766.2 and NC_013768.1. The other 34 genomes are expected to be released on GenBank soon. These genomes are all from clinical isolates from across Canada, and were isolated over many years ranging from 1981 to 2010. They were sequenced using the Illumina MiSeq platform, with 250 bp paired end reads and an insert size of 0 bp. Complete genomes were obtained by sequencing select regions with Sanger sequencing in order to bridge gaps between contigs.

Each draft genome was assembled using SPAdes (Bankevich et al., 2012), which was chosen because multiple assembly evaluations determined that SPAdes had the best performance with bacterial genomes (see section 2.1.1). SPAdes produced high quality draft genome sequences. The average number of contigs in these draft genomes was 66 (the range was 23-231 contigs). The average number of contigs greater than 1000 bp, which is considered a better reflection of genome quality than the total number of contigs, was 14 (the range was 10-24 contigs). The average N50 was 534 kb (the range was 297 kb to 1.5 Mb). The average percentage of CDSes missing from the draft genomes was 0.65%, but it should be noted that most draft genomes were missing between 0.097% and 0.61% of their total CDSes. A single draft genome, 0861, was missing 10% of CDSes. This genome had an N50 of 561125 bp and contained 12 contigs greater than 1000 bp, but not all of these contigs could be aligned to the complete genome. This may be due to errors during sequence assembly. This is an example of how N50 and number of contigs are imperfect measures of genome quality.

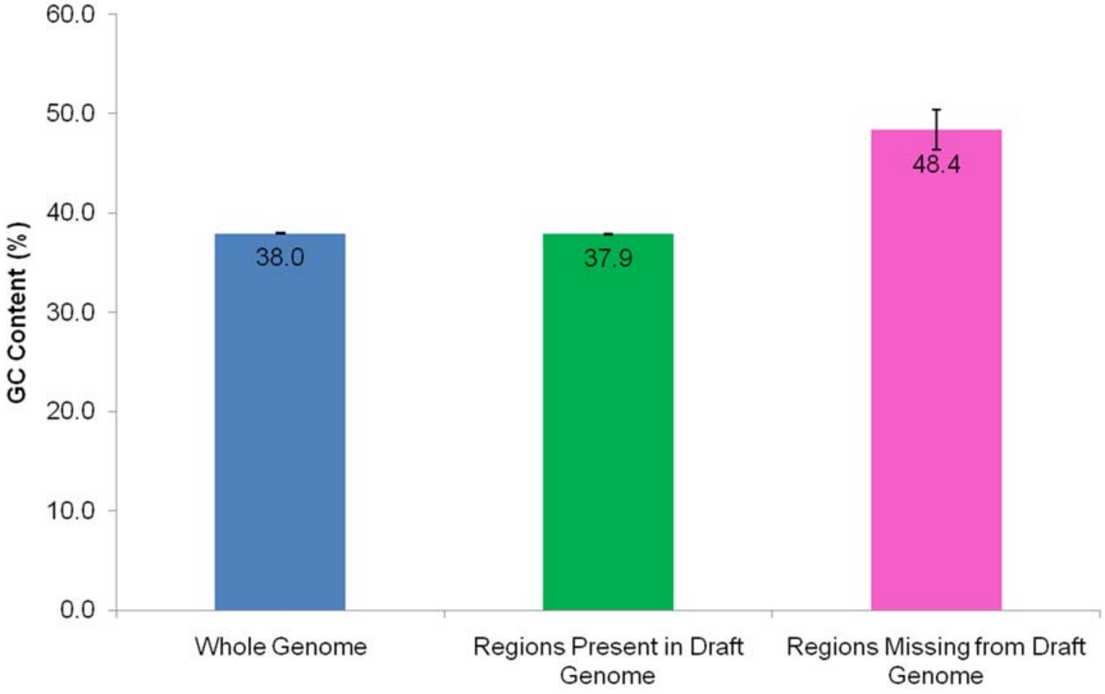
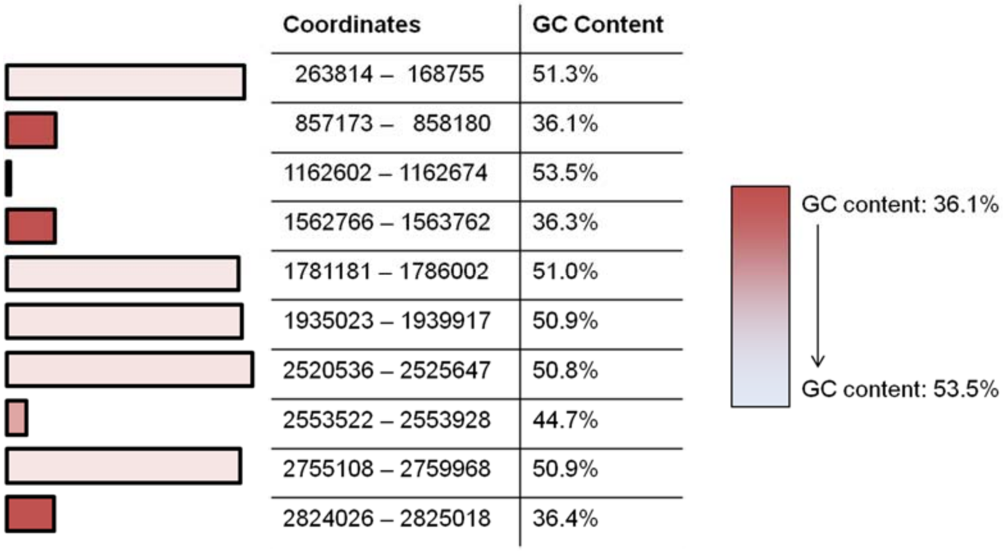


Figure 2.1 Average GC content of the 36 *Listeria monocytogenes* complete genomes, draft genomes, and regions missing from draft genomes. Standard deviation in GC content of each region type is shown.

The average genome size for this data set was 2.98 ± 0.04 Mb. The average GC content was 37.97 ± 0.02%. GC content of the draft genomes were similar but slightly lower, with an average of 37.86 ± 0.02%. Regions missing from the draft genomes, however, had a higher GC content at 48 ± 2%. A comparison of these values is shown in figure 2.1. While there was little variation in GC content of missing regions across genomes, there was more variation in GC content of the individual missing regions within a single draft genome. Figure 2.2 shows the GC content of the regions missing from genome 95-0093.



Coordinates	GC Content
263814 – 168755	51.3%
857173 – 858180	36.1%
1162602 – 1162674	53.5%
1562766 – 1563762	36.3%
1781181 – 1786002	51.0%
1935023 – 1939917	50.9%
2520536 – 2525647	50.8%
2553522 – 2553928	44.7%
2755108 – 2759968	50.9%
2824026 – 2825018	36.4%

Figure 2.2 Regions missing from a single *Listeria monocytogenes* draft genome (95-0093) and the GC content of those regions.

In order to simulate a project where only draft genomes are available, complete genomes from this set of 36 isolates were not used as reference genomes in IslandViewer (more details about IslandViewer analysis are described in section 2.5). Reference genomes for IslandViewer analysis were chosen by performing a MegaBLAST alignment of the largest contig from each draft genome against all complete bacterial genomes in GenBank, excluding genomes from this data set. Table 2.1 shows the reference genomes used for IslandViewer analysis.

Table 2.1 *Listeria monocytogenes* genomes used as a reference for IslandViewer analysis

Listeria Genome Name	Similar Reference
95-0093	NC_018593.1
88-0478	NC_018593.1
81-0558	NC_021825.1
81-0592	NC_021825.1
10-0809	NC_021825.1
10-0810	NC_021824.1
10-0811	NC_021824.1
10-0812	NC_017544.1
10-0813	NC_017544.1
10-0814	NC_018593.1

10-0815	NC_021837.1
81-0861	NC_021825.1
10-0933	NC_021837.1
10-0934	NC_021837.1
10-1046	NC_018593.1
10-1047	NC_018593.1
02-1103	NC_019556.1
02-1289	NC_019556.1
10-1321	NC_021837.1
02-1792	NC_019556.1
98-2035	NC_018593.1
10-4754	NC_021837.1
10-5024	NC_018593.1
10-5025	NC_018588.1
10-5026	NC_018588.1
10-5027	NC_018588.1
04-5457	NC_018593.1
02-5993	NC_018593.1
08-6056	NC_021837.1
99-6370	NC_018593.1
08-6569	NC_021837.1
02-6679	NC_021824.1
02-6680	NC_021824.1
08-6997	NC_021837.1
08-7374	NC_018593.1
08-7669	NC_018593.1

2.3.2. Genes of interest

AMR gene homologs and VF orthologs were detected in each of the 36 complete *L. monocytogenes* genomes, and 11 ± 6 AMR genes were predicted per complete genome. For all but one of the genomes, every AMR gene was present in the draft genome. A single draft genome, *L. monocytogenes* 02-1792, was missing one of the 18 genes detected in the complete genome. The missing gene was predicted to encode an efflux pump that confers resistance to macrolides. This family of efflux pumps was first discovered in *Streptomyces* and belongs to a larger family of ATP-dependent transport proteins (Schoner et al., 1992). For each of the 36 genomes, the proportion of missing AMR genes was lower than the proportion of all CDSes missing from the draft. However, the difference between these proportions is not statistically significant based on two tailed Z-test. The complete genomes contained 64 ± 6 predicted VFs per genome. Every VF was present in the draft version of each genome, so as with predicted AMR genes, the proportion of missing VFs was lower than the proportion of all CDSes missing from the draft for each of the 36 genomes. Similar to AMR genes, the difference between these proportions is not statistically significant different based on a two tailed Z-test. Figure 2.3 shows a box plot representation of the AMR genes and VFs missing from the *Listeria* draft genomes.

tRNA genes were also assessed as a gene of interest because of their association with MGEs (see section 1.2). Each of the complete *Listeria* genomes has exactly 58 tRNA genes in their PROKKA annotations, and all except one of the draft *Listeria* genomes was missing 4 of those tRNA genes, or 6.9%. The genomic locations and annotations of the missing tRNA genes in each genome were assessed manually, and different sets of 4 tRNA genes were missing from different draft genomes. The single exception is *L. monocytogenes* 81-0861, which is also the genome whose draft version is missing a larger percentage of total CDSes. The draft version of this genome was missing 24 tRNA genes, or 41.4%. This genome brought up the average percentage missing to 7.7%. When compared to the percentage of CDSes missing from the draft versions of these genomes, the difference is not only statistically significant as a whole, but it is statistically significant in each individual genome ($p < 10^{-5}$).

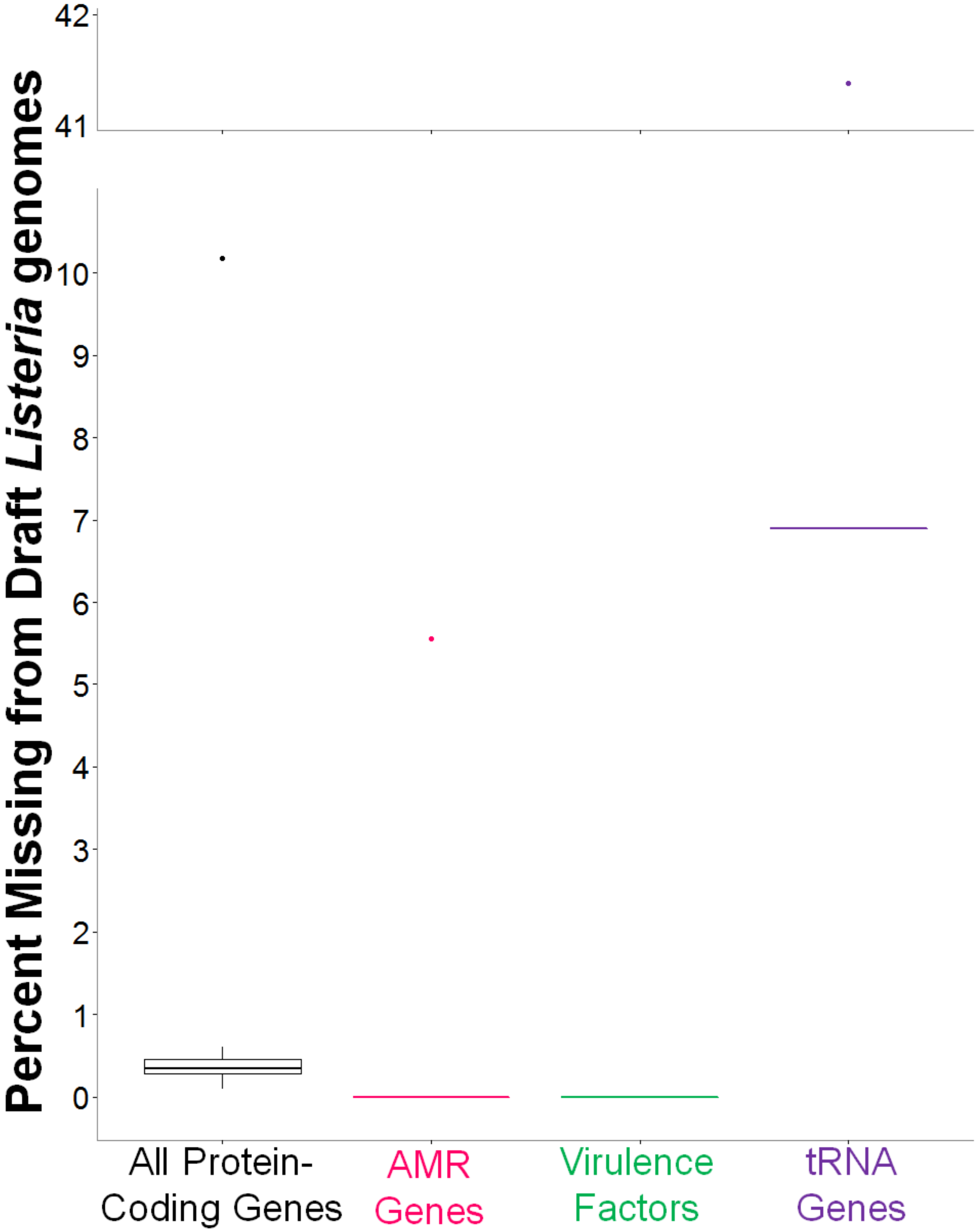


Figure 2.3 **Box plot of the percentage of genes in a complete genome that are missing from the draft version of the same genome. A single outlier is shown on a separate plot above the main box plot.**

2.3.3. COG analysis

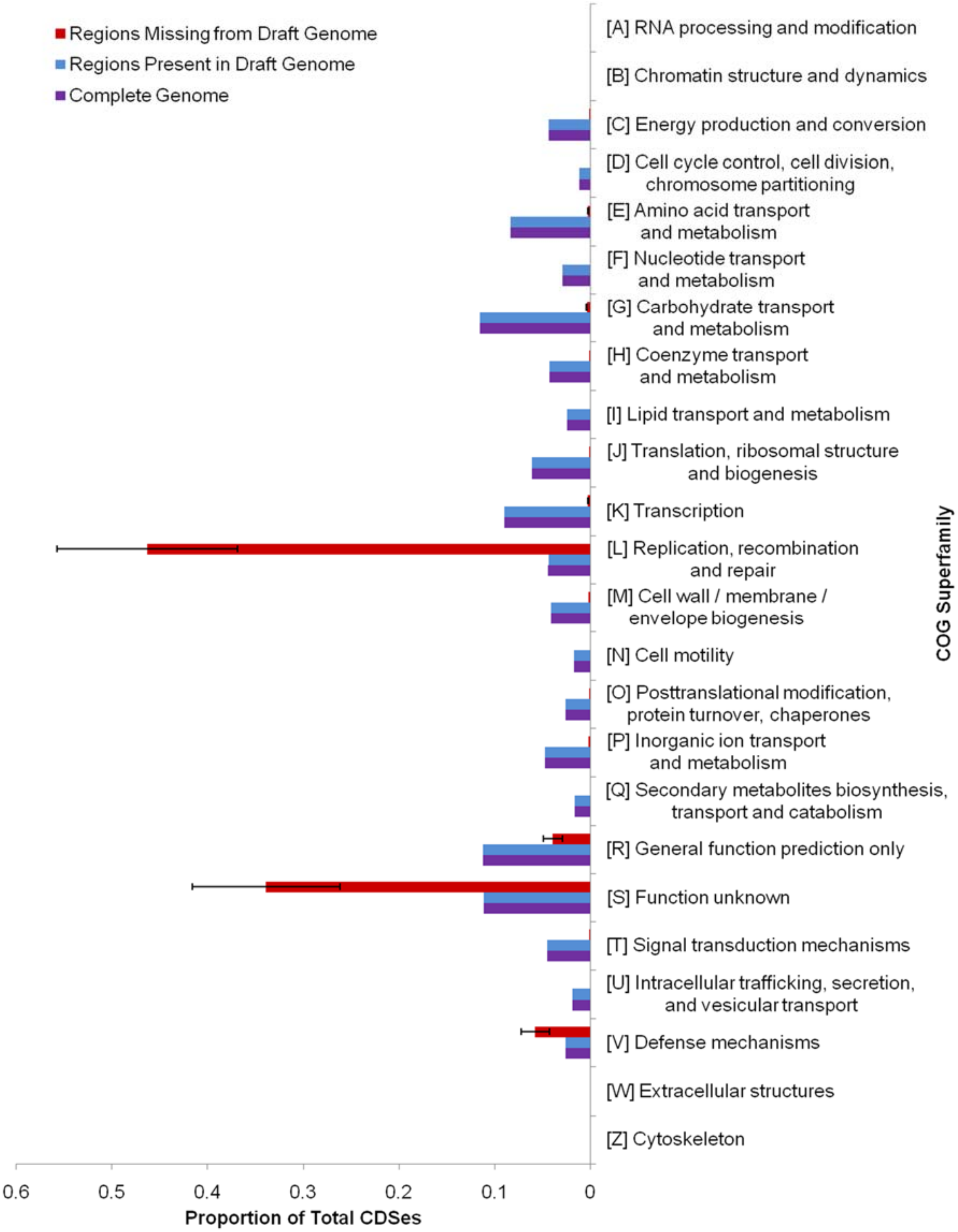


Figure 2.4 Distribution of COG superfamilies amongst CDSes in complete *Listeria monocytogenes* genomes, and amongst CDSes that are present or missing from the draft versions of these genomes.

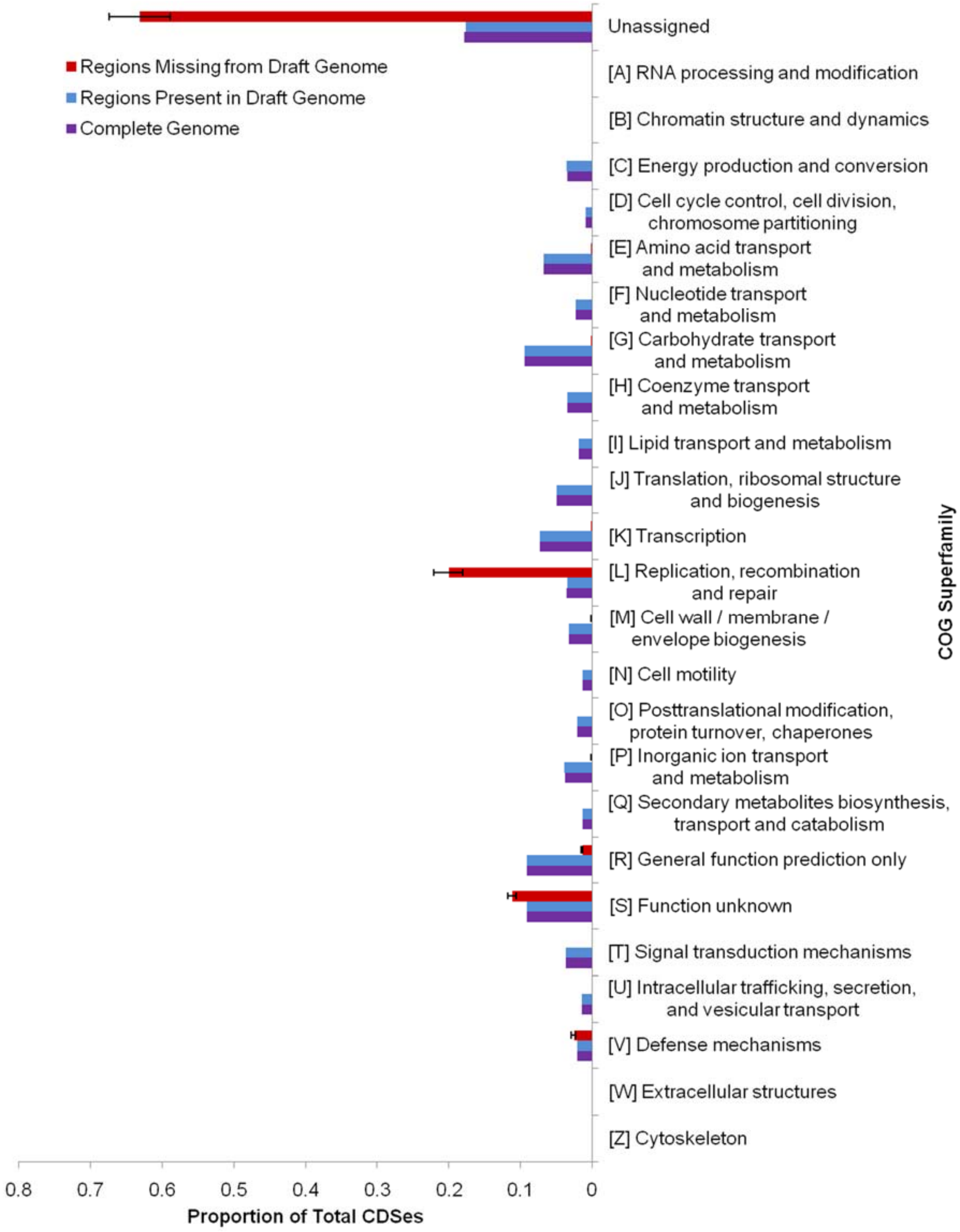


Figure 2.5 Distribution of COG superfamilies amongst CDSes in complete ***Listeria monocytogenes*** genomes, and amongst CDSes that are present or missing from the draft versions of these genomes. This figure includes CDSes which could not be assigned to a COG category.

COG categories were assigned to CDSes in the 36 complete *L. monocytogenes* genomes, and 82.1 ± 9.7% of CDSes were able to be assigned to a COG category. The distribution of COG superfamilies among the assigned CDSes is shown in purple in Figure 2.4. The distribution of COG superfamilies were also measured in regions missing from the draft version of *Listeria* genomes to determine whether there is a bias in COG categories in these missing regions. If there was no bias in gene coverage, the distribution of proportion would be expected to be equivalent. The distribution of COG superfamilies among assigned CDSes in missing regions is shown in red in figure 2.4. This figure is meant to be analogous to a previously produced figure from another group showing the distribution of COG superfamilies amongst partially covered ORFs (Klassen & Currie, 2012). In that case, replication, recombination and repair, signal transduction mechanisms, and secondary metabolites biosynthesis were the overrepresented superfamilies among partially covered ORFs in 25 draft *Streptomyces* genomes.

In order to show standard deviation, the proportions in the Figure 2.4 are the means of the proportions for the 36 *Listeria* genomes. Using an average proportion across 36 genomes potentially creates a bias towards genes missing from higher quality genomes; if a draft genome is missing 10 genes, each missing gene will affect the average proportion 10-fold more than a gene in a draft genome which is missing 100 genes. To avoid this bias, proportions that were used to test for statistical significance were calculated from the sum of genes from each of the 36 *Listeria* genomes for each superfamily.

Because a large proportion of CDSes in missing regions were not assigned to a COG category, superfamilies which are overrepresented amongst assigned CDSes may not be overrepresented amongst CDSes as a whole. Figure 2.5 shows the distribution of CDSes in complete genomes and regions missing from draft genomes. Unlike Figure 2.4, the distributions in Figure 2.5 include CDSes which were not assigned to any COG category. Being overrepresented only amongst assigned CDSes may still be meaningful; the distribution of gene functions amongst unassigned CDSes is unknown, and could mirror the distribution of genes with predicted functions. Metagenomics studies have compared distributions of assigned COGs across data sets (Gosalbes et al., 2011; Kurokawa et al., 2007), although in the case of a metagenomics study, different proportions of unassigned COGs could be due to uncharacterized species, and this explanation does not apply here.

Using Fisher's exact test, replication, recombination, and repair was the only superfamily which was overrepresented in regions missing from the complete genomes ($p = 4.76 \times 10^{-18}$). This superfamily was significantly overrepresented even after a Bonferroni correction for multiple sample testing. The majority of CDSes missing from this superfamily were assigned to the Transposase and inactivated derivatives category, COG2801 (69/80), and the vast majority of genes assigned to this COG category were missing from draft genomes in this data set (69/71).

Unassigned CDSes were also significantly overrepresented in regions missing from draft genomes even after a Bonferroni correction (p = 4.89 x10-66). In total, 1.6% of unassigned CDSes were missing from the set of draft *L. monocytogenes* genomes, and 63 ± 4% of all missing CDSes were unassigned. The majority of these unassigned CDSes were annotated as conserved hypothetical proteins by PROKKA (173/297). A similar analysis of COG categories that were overrepresented in GI regions has previously been performed (Hsiao et al., 2005). Notably, CDSes which could not be assigned to a COG category and those assigned to the replication, recombination, and repair superfamily were also found to be the two overrepresented groups in GIs.

2.4. Draft *Pseudomonas aeruginosa* genomes

2.4.1. The data set

The *P. aeruginosa* data set for this analysis is from the *Pseudomonas aeruginosa* reference panel. This panel was developed in order to represent the diversity of the species, and has a mix of older, commonly studied isolates and more recently isolated strains. The panel consists of 40 *P. aeruginosa* strains isolated around the world from both clinical and environmental samples (De Soyza et al., 2013). The genomes of these isolates were produced using various sequencing technologies and assembly algorithms, and seven of the genomes are complete. Draft genome data is available for two complete genomes, both of which were sequenced using pyrosequencing technology and completed by Sanger sequencing of genomic regions between contigs (Jeukens et al., 2014).

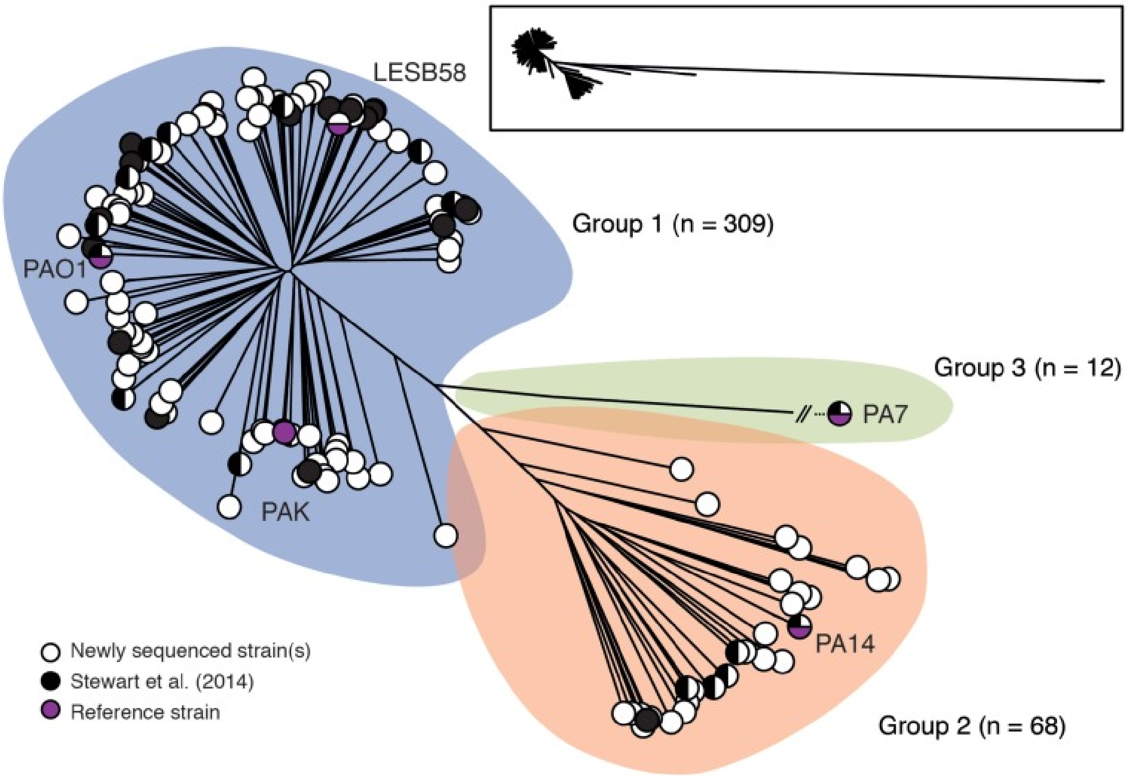


Figure 2.6 Phylogenetic tree that was used to identify complete genomes that are closely related to draft panel genomes. Figure is from Freschi *et al*, 2015 (Freschi et al., 2015) **and is licensed under CC BY.**

To increase the sample size of matching draft and complete genomes, a phylogenetic tree of *P. aeruginosa* genomes produced using parSNP (Treangen et al., 2014) was parsed to find the closest complete genome to each draft genome in the reference panel. Based on the distribution of distances from panel genomes to the nearest complete genomes, the maximum distance between a draft genome and the complete genome used for comparative analysis was set to 0.01. Due to the timing of the analysis, some complete genomes were not available to use as a reference for IslandViewer analysis (Dhillon et al., 2015), so a similar search of the phylogenetic tree was done to identify the most similar complete genome available to use as a reference in IslandViewer. In total, 11 of the 40 panel genomes met the requirements for this analysis, including the two complete genomes for which draft data is still available.

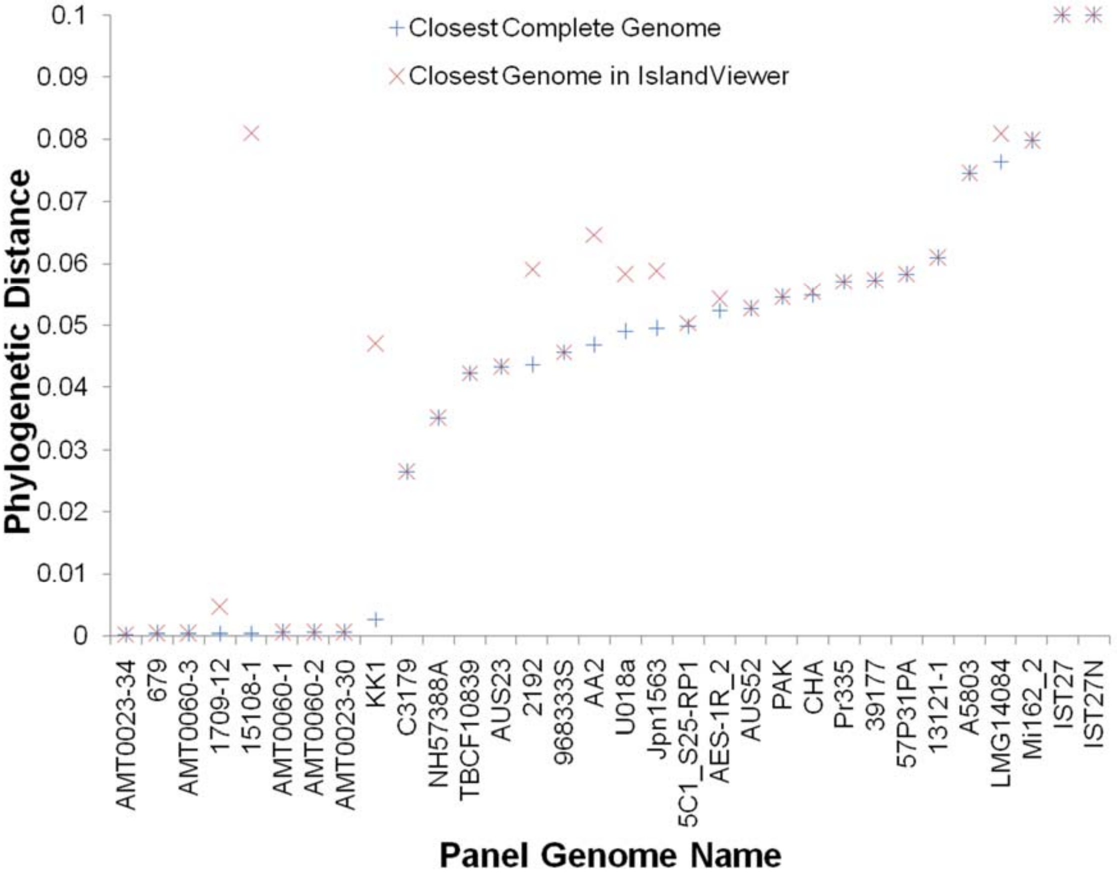


Figure 2.7 Graph showing distances of *Pseudomonas aeruginosa* draft panel genomes from the nearest complete genome. CPHL9433, which had a distance of 1.8353 from any complete genome, was excluded from this figure for clarity. Genomes are listed in ascending order of distance from the closest complete genome.

The sources of the 11 genomes used for this analysis include seven isolates from cystic fibrosis patients, one from the parent of a cystic fibrosis patient, and three other clinical, non-cystic fibrosis samples. The nine genomes for which there is no complete genome from the same isolate were all sequenced using Illumina MiSeq, with a TruSeq paired-end 2x300 library. The other two genomes, which were sequenced earlier and have since been completed were sequenced using 454 GS-FLX Titanium (Jeukens et al., 2014). The N50 of all 11 genomes used ranges from 200 kb to 479 kb. The percentage of CDSes absent from these genomes but present in the most similar complete genome ranged from 0.26% to 9.27%, with an average of 2%.

This analysis is part of a much larger project to analyze the genomes of the *Pseudomonas aeruginosa* panel strains. Careful analyses of these 40 strains, which represent the diversity of the species, will lead to a better understanding of *P. aeruginosa* evolution. The assessment of draft genomes described here is meant to provide support for the use of draft genomes in the broader panel genome project.

Table 2.2 *Pseudomonas aeruginosa* genomes used for draft analysis

Panel Genome	Complete Reference	Complete Reference for IslandViewer Analysis
15108-1	NZ_CP011369.1	NC_017549.1
1709-12	NZ_CP011317.1	NZ_CP010555.1
679	NC_021577.1	NC_021577.1
AMT0023-30	NZ_AAQW01000001.1	NZ_AAQW01000001.1
AMT0023-34	NZ_AAQW01000001.1	NZ_AAQW01000001.1
AMT0060-1	NZ_CP010555.1	NZ_CP010555.1
AMT0060-2	NZ_CP010555.1	NZ_CP010555.1
AMT0060-3	NZ_CP010555.1	NZ_CP010555.1
KK1	NZ_CP008749.1	NZ_CP007147.1
LES400	NZ_CP006982.1 (same isolate)	NC_023066.1
LES431	NC_023066.1 (same isolate)	NC_023066.1 (same isolate)

A significant limitation to the *P. aeruginosa* data set in comparison to the *L. monocytogenes* data set is that the majority of the draft genomes do not have complete genomes from the same isolate to be used as a reference. This means that the results may be impacted by actual differences between isolates. The analysis is restricted to draft genomes and complete genomes from very closely related isolates in order to reduce the impact of this limitation. However, in section 2.3, and again in this section, it is found that regions associated with mobile elements are more prevalent in regions missing from draft genomes. Since differences in mobile elements can be present even between strains that are very closely related in a core genome phylogenetic tree (Hao & Golding, 2006), not having complete genome sequences from the same isolates to use as references greatly limits the strength of this study in the *P. aeruginosa* data set. For this reason, the *P. aeruginosa* data set should be considered a secondary, supporting data set to the main analysis that was performed using *L. monocytogenes* data.

2.4.2. Genes of interest

AMR gene homologs and VF orthologs were detected in the 8 complete *P. aeruginosa* genomes listed in the middle column of table 2.2, and the coordinates of these homologs were compared with regions missing from a draft genome of a similar or identical isolate, which is listed in the left column of table 2.2. The complete genomes contained 46 ± 5 predicted AMR genes per genome. Five draft genomes were missing one AMR gene each, while the other six draft genomes contained all of the AMR genes found in their most similar complete genome, bringing the average percentage of AMR genes missing from drafts to 1 ± 1%. The complete genomes contained 150 ± 40 VF orthologs per genome. The percentage of VFs that were missing from draft genomes was also 1 ± 1%. The proportion of AMR genes was not significantly different from the total proportion of CDSes, but the proportion of VFs missing from draft *Pseudomonas* genomes was significantly lower than the total proportion of missing CDSes ($P < 10^{-5}$). Figure 2.8 shows a box plot representation of the AMR genes and VFs missing from the *Pseudomonas* draft genomes.

Unlike the *Listeria* data set, tRNA gene annotations for the *Pseudomonas* data set were extracted from RefSeq annotations which use the NCBI Prokaryotic Genome Annotation Pipeline including tRNAscan-SE for tRNA predictions (Lowe & Eddy, 1997). The prediction method used is for tRNA genes is not expected to a significant impact on the results. As with the *Listeria* data set, tRNA genes were significantly overrepresented in regions that were not present in draft genomes. Complete *Pseudomonas* genome annotations contained 69 ± 6 tRNA genes per genome, and 11 ± 3% of these genes were in regions that were not present in the draft genomes. When compared to the percentage of CDSes missing from the draft genomes, tRNA genes are strongly overrepresented in missing regions ($P < 10^{-5}$).

It should be noted that multiple draft genomes are being compared to a single complete genome for this data set, and complete genomes which are used as a reference for more than one draft may have a larger impact on the results (see Table 2.2).

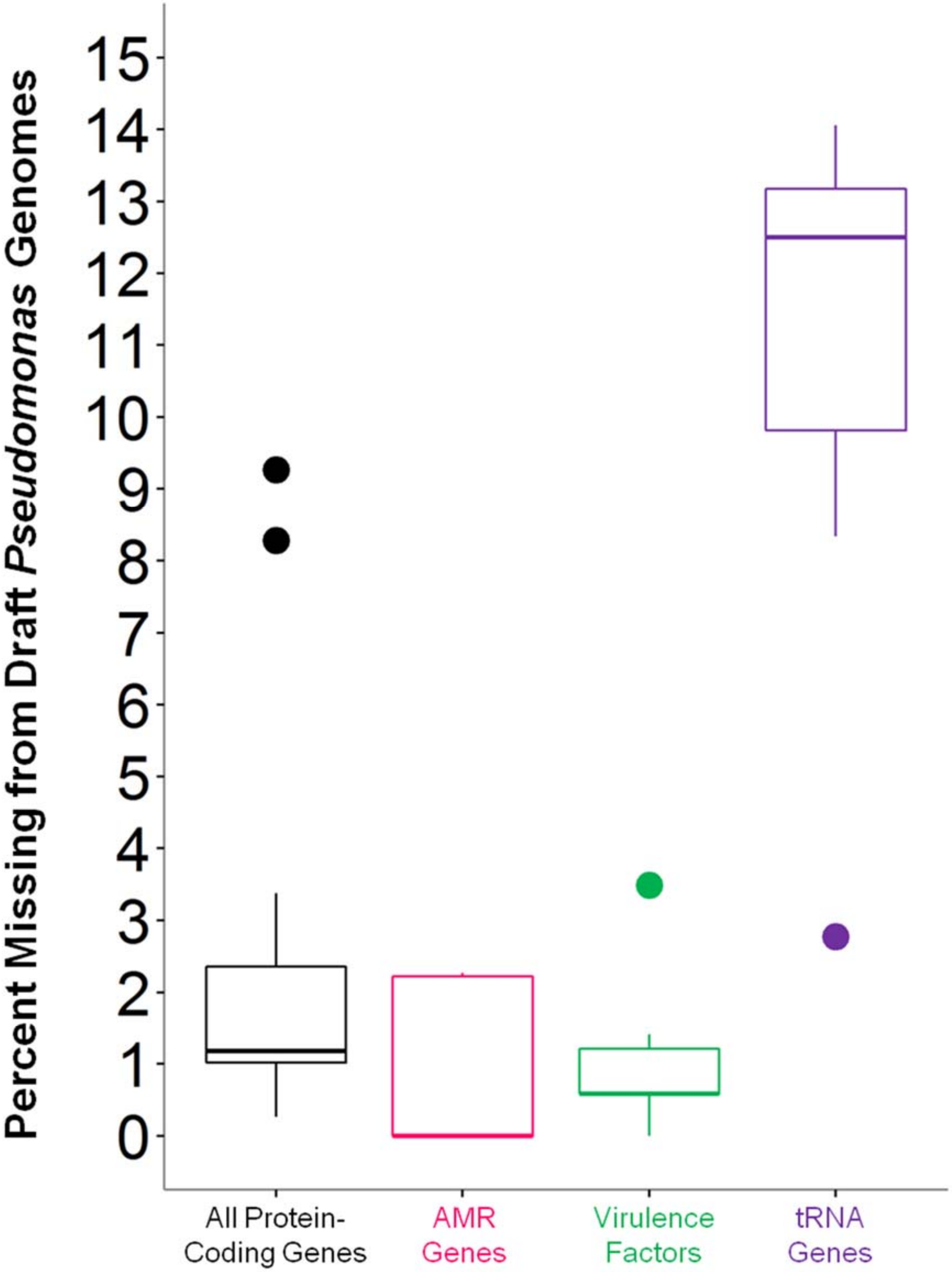


Figure 2.8 **Box plot showing, for several gene types, the percentage of genes that are missing from draft *Pseudomonas aeruginosa* genomes.**

2.4.3. COG analysis

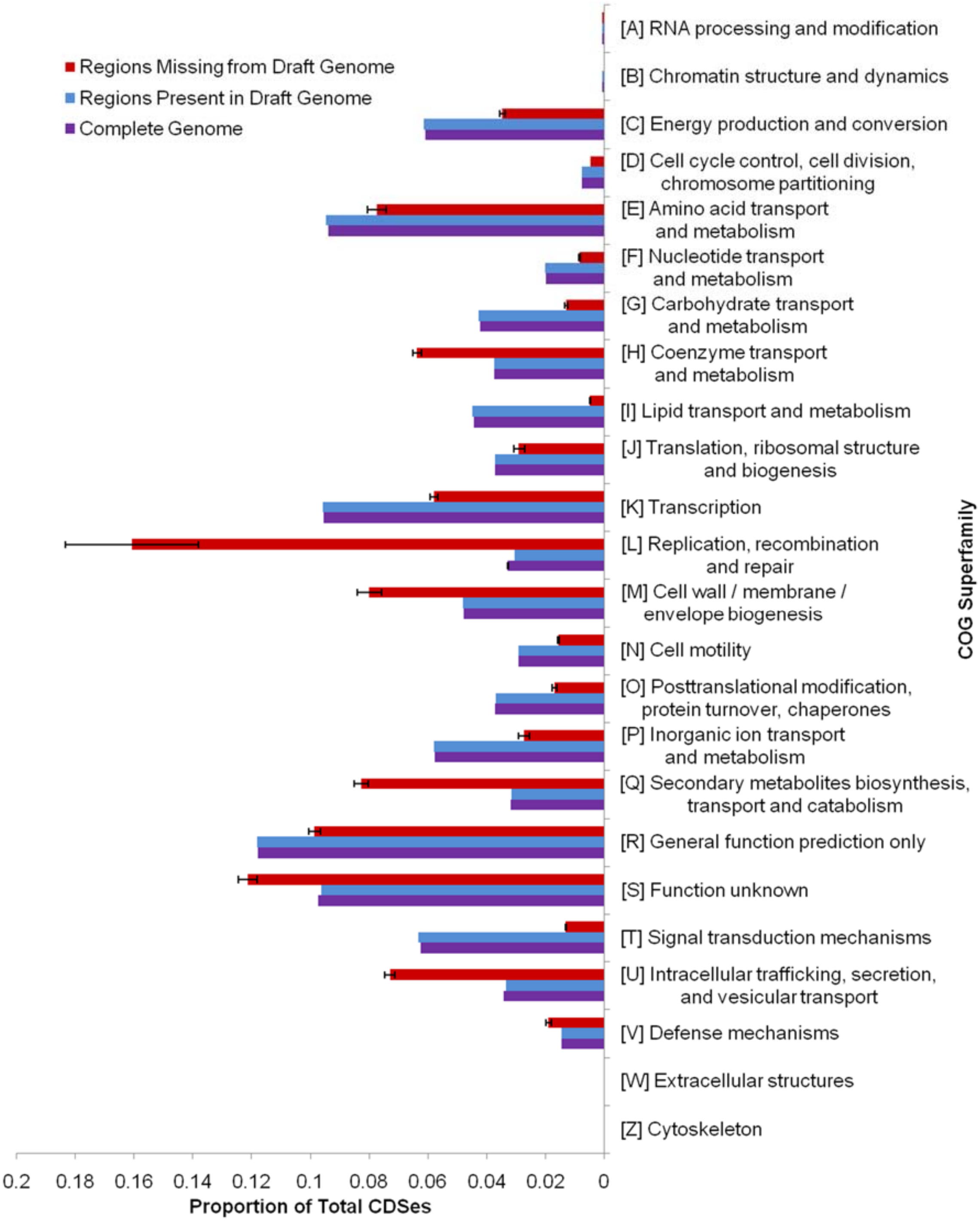


Figure 2.9 Distribution of COG superfamilies amongst CDSes in complete *Pseudomonas aeruginosa* genomes, and amongst CDSes that are present or missing from draft genomes of very similar or identical isolates.

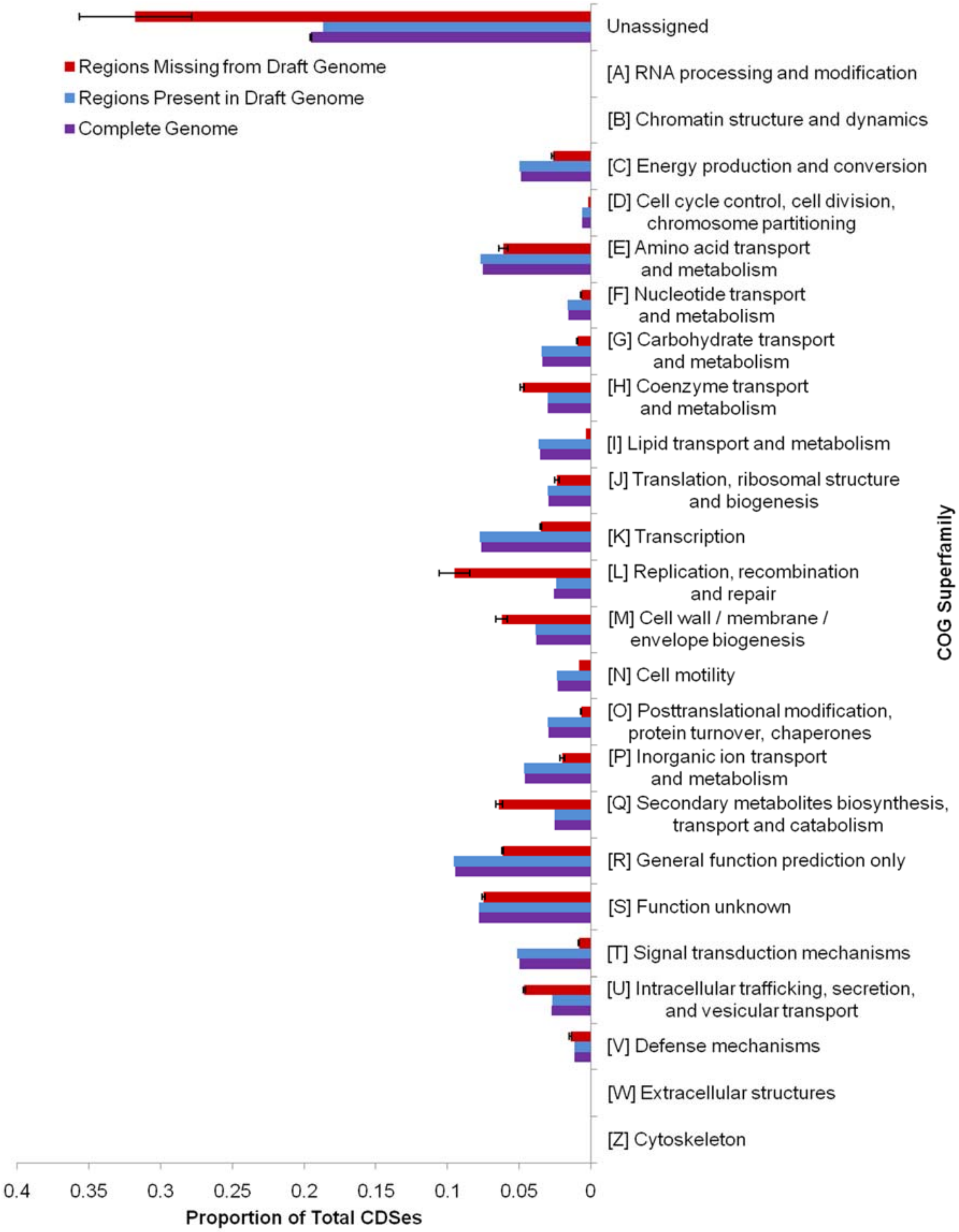


Figure 2.10 Distribution of COG superfamilies amongst CDSes in complete *Pseudomonas aeruginosa* genomes, and amongst CDSes that are present or missing from draft genomes of very similar or identical isolates. This figure includes CDSes which could not be assigned to a COG category.

COG categories were assigned to CDSes in the 8 complete *P. aeruginosa* genomes listed in the middle column of table 2.2, and 80 ± 2% of CDSes were able to be assigned to a COG category. The distribution of COG superfamilies among the assigned CDSes is shown in purple in Figure 2.9, but it should be noted that because this is meant to be compared to regions missing from the draft *Pseudomonas* genomes, NZ_AAQW01000001.1 was counted twice and NZ_CP010555.1 was counted three times towards towards the average and standard deviation values in the figure (see table 2.2). The distribution of COG superfamilies among assigned CDSes in regions missing from draft *Pseudomonas* genomes is shown in red in Figure 2.9.

As with the *Listeria* data set, the only COG superfamily that was significantly overrepresented in missing regions was the replication, recombination, and repair superfamily (p=7.94e-29). CDSes which could not be assigned to a COG category were also overrepresented in regions missing from draft genomes (p = 2.32e-180). Significance was tested using Fisher's exact test, similarly to the *L. monocytogenes* data set as described in section 2.3.3. They were both still significantly overrepresented after the Bonferroni correction for multiple sample testing. In total, 8% of unassigned CDSes were missing from the set of draft *Pseudomonas* genomes compared to the most similar complete genomes, and 30 ± 20% of all missing CDSes were unassigned for each genome.

As with the *Listeria* data set, because a large proportion of CDSes in missing regions were not assigned to a COG category, superfamilies which are overrepresented amongst assigned CDSes may not be overrepresented amongst CDSes as a whole. The distribution of COG superfamilies among all CDSes, including those which could not be assigned to a COG category, is shown in Figure 2.10. The cell wall, membrane, and envelope biogenesis and intracellular trafficking, secretion, and vesicular transport superfamilies represent a higher proportion of assigned CDSes in missing regions, as shown in figure 2.9, but were not found to be significantly overrepresented using the measurement of CDSes described in 2.3.3. Some superfamilies are, in fact, significantly underrepresented.

2.5. Genomic islands and contig boundaries

2.5.1. Method overview

To assess GI predictions in draft genomes, the GI predictions in draft genomes were compared with GI predictions in complete genomes. The flowchart below outlines the information used for the comparison. Draft genome GI predictions were mapped to coordinates in the complete genome using MegaBLAST, using requirements similar to those used to align contigs to the complete genome (this was described in section 2.2.1), but with less strict limits on alignment length variation. For GIs, the alignment had to cover a minimum of 75% of the GI sequence length, and have gaps of no more than the length of the GI sequence (these limits on alignment length and identity were chosen empirically through manual assessment of a range of limits).

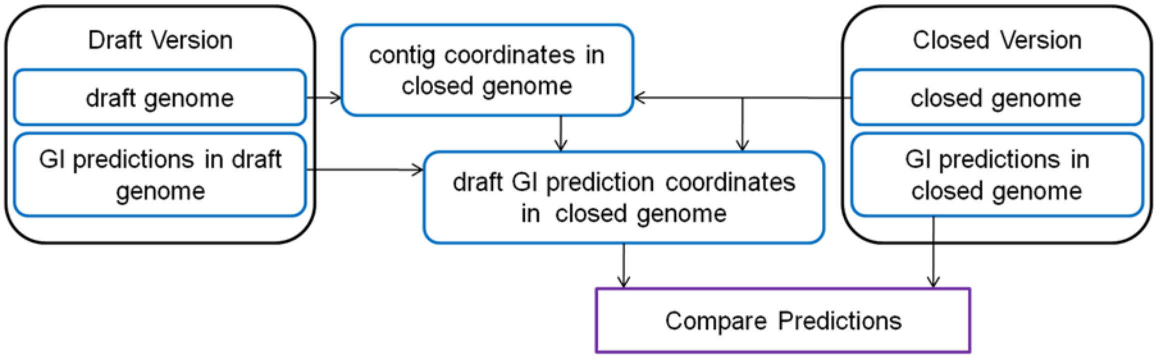

Draft Version
draft genome
GI predictions in draft genome
contig coordinates in closed genome
draft GI prediction coordinates in closed genome
Closed Version
closed genome
GI predictions in closed genome
Compare Predictions

Figure 2.11 Flowchart illustrating basic pipeline for assessment of draft genomes with IslandViewer

True positives, false positives, and false negative (TP, FP, FN) GI predictions were calculated in two different ways: a 50% overlap method, and a per-base method. For the 50% overlap method, each GI that is predicted in both the draft and complete genome is counted as a TP, and a GI is considered to be the same if the prediction in the draft genome and complete genome overlap by at least 50% of their lengths. FPs are GIs predicted only in the draft genome, and FNs are genomic islands predicted only in the complete genome, and these are both determined by having less than a 50% overlap with GIs in the other version of the equivalent genome. For the per base method, each base in the genome (as opposed to each GI) may be a TP, FP, FN, or TN (true negative). Each base that is predicted to be a GI in both the complete and draft genome is a TP. Bases in the complete genome that are predicted to be an island but are not predicted in the draft genome are FNs, and bases that are predicted to be GIs in the draft genome but are not predicted to be GIs in the complete genome are FNs. Bases that are predicted to be GIs in neither the draft nor complete version of a genome are TNs. After using either method to calculate TP, FP, and FN, precision and sensitivity were calculated using the equations below.

precision = TP / (TP + FP)

sensitivity = TP / (TP + FN)

accuracy = (TP + TN) / (TP + TN + FP + FN)

It is important to note that IslandViewer GI predictions in the complete genome are not equivalent to the true set of GIs in each genome. Like every GI prediction method, the methods in IslandViewer may make incorrect predictions or fail to predict true islands, even when using a complete genome as input. This is true even though IslandViewer uses three of the most accurate GI prediction methods (Dhillon et al., 2015). See section 1.3 for more information about GI prediction methods. For the purposes of this analysis, IslandViewer predictions are being treated as TPs for two reasons. First, it is outside the scope of this thesis to experimentally verify the locations of true GIs (such as activation of bacteriophages by exposing cells to mitomycin (van Schaik et al., 2010)) within the genomes of each isolate in the *Listeria* and *Pseudomonas* data sets. Second, as this is an assessment of the utility of draft genomes vs. complete genomes, it is more straightforward to compare IslandViewer predictions in draft genomes vs. IslandViewer predictions in complete genomes, rather than comparing IslandViewer predictions in both draft and complete genomes vs. experimentally verified GIs.

2.5.2. GenomeD3Plot visualizations

GenomeD3Plot is a javascript library built using the D3 library (Laird et al., 2015). Originally developed for IslandViewer (the incorporation of GenomeD3Plot into IslandViewer is described in more detail in section 3.5), GenomeD3Plot is a flexible tool for producing genome visualizations that incorporate various genome features such as genomic islands, gene content, and GC content. GenomeD3Plot was used to generate visualizations of contigs aligned to complete genomes, overlaid with GI predictions for the complete genome and equivalent draft genome.

Two types of GenomeD3Plot JSON scripts were automatically generated for each draft-complete genome pair, and these scripts were used to produce two distinct visualizations. The first, more complex, visualization displays GIs predicted by each method in IslandViewer, and displays the alignment of each contig to the complete genome. The second, simplified visualization displays integrated GI predictions and regions missing from the draft genome.

Two example genomes are shown here: *L. monocytogenes* 95-0093 and *P. aeruginosa* LES431. In the *Pseudomonas* example, an IslandPath-DIMOB prediction in the draft genome with the coordinates 4878834-4910304bp is shown as not matching in the more complex visualization that separates predictions by method, but because SIGI-HMM predicts the GI at that location in the complete genome, the GI is labelled as matching in the simplified visualization. This is an example of the benefits of using multiple GI prediction tools, as multiple tools are able to complement each other (Langille et al., 2008a). Both examples have locations where GI predictions overlap with regions missing from the draft genome. The plots are useful for quickly interpreting the relative locations of GIs and contig breaks, and possible relationships between the two. This relationship was assessed quantitatively as well, and this assessment will be discussed in the following two sections.

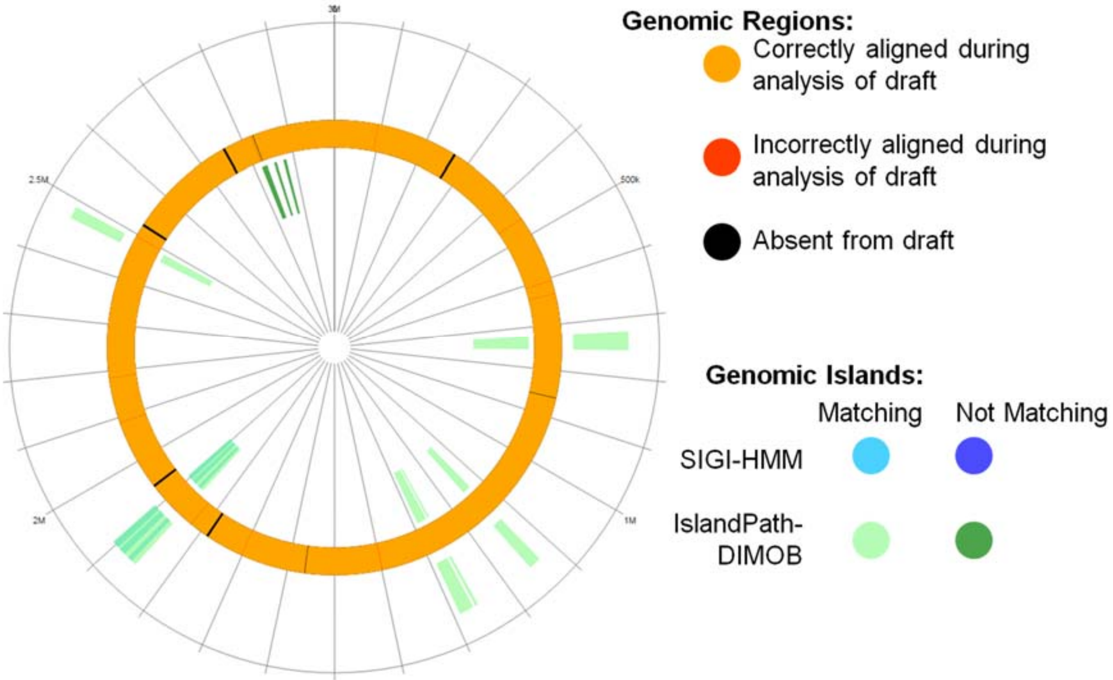


Figure 2.12 Example: visualization of GI predictions in *Listeria monocytogenes* 95-0093. Draft predictions are displayed on the outer ring, and predictions in the complete genome are displayed on the inner ring. Matching GI predictions were predicted in both the draft and complete genome. When SIGI-HMM and IslandPath-DIMOB predictions overlap, it produces a cyan colour. GI predictions on or near regions absent from the draft were not predicted in the draft genome.

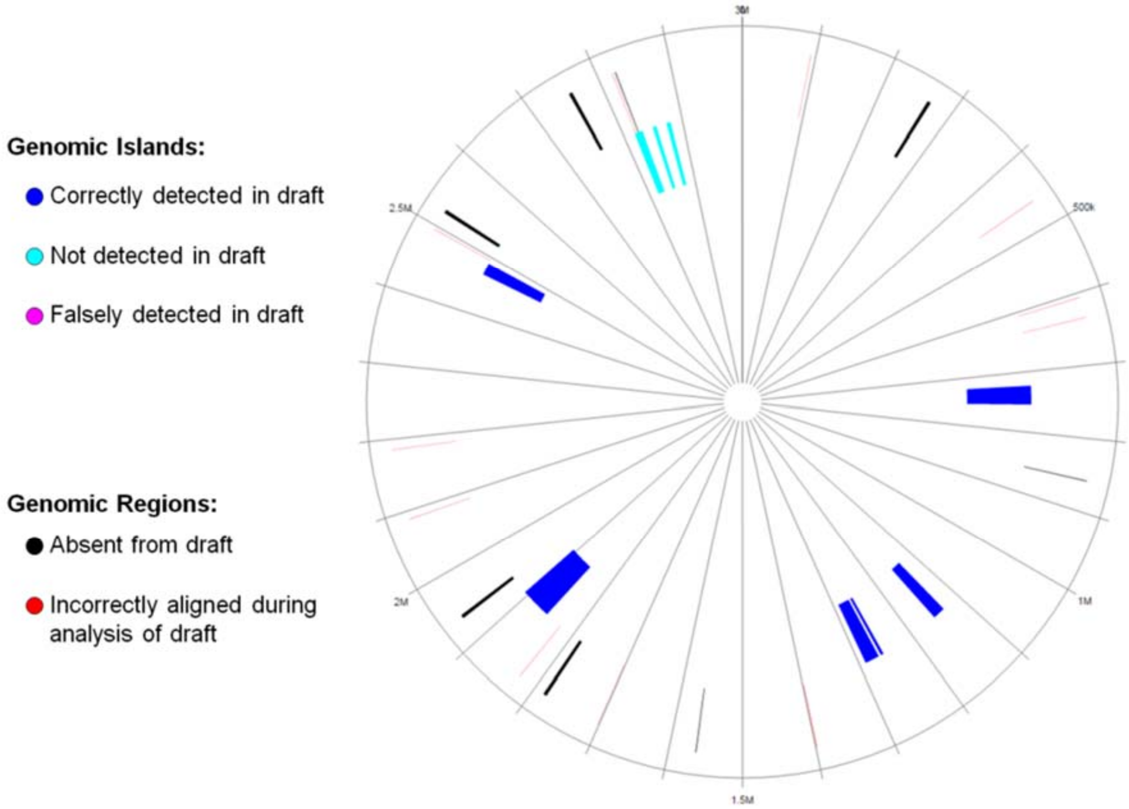


Figure 2.13 Example: simpler visualization of GI predictions in *Listeria monocytogenes* 95-0093.

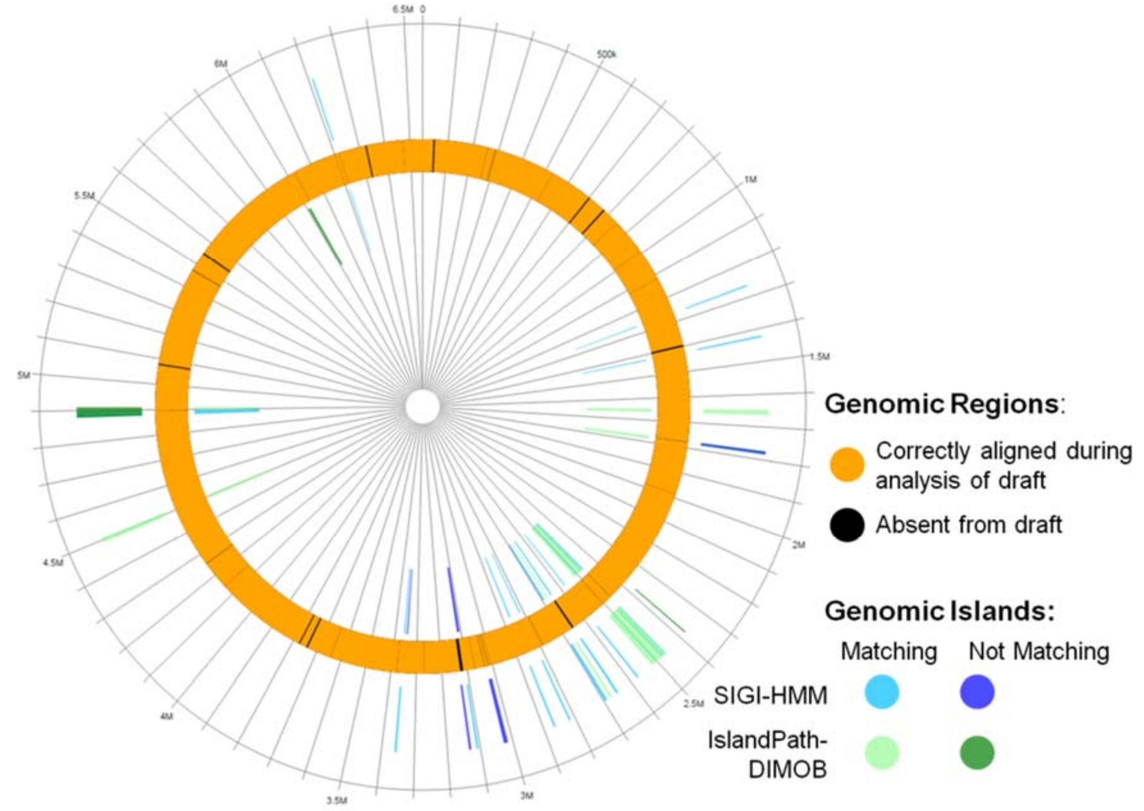


Figure 2.14 **Example: visualization of GI predictions *Pseudomonas aeruginosa* LES431. Draft predictions are displayed on the outer ring, and predictions in the complete genome are displayed on the inner ring. Matching GI predictions were predicted in both the draft and complete genome.**

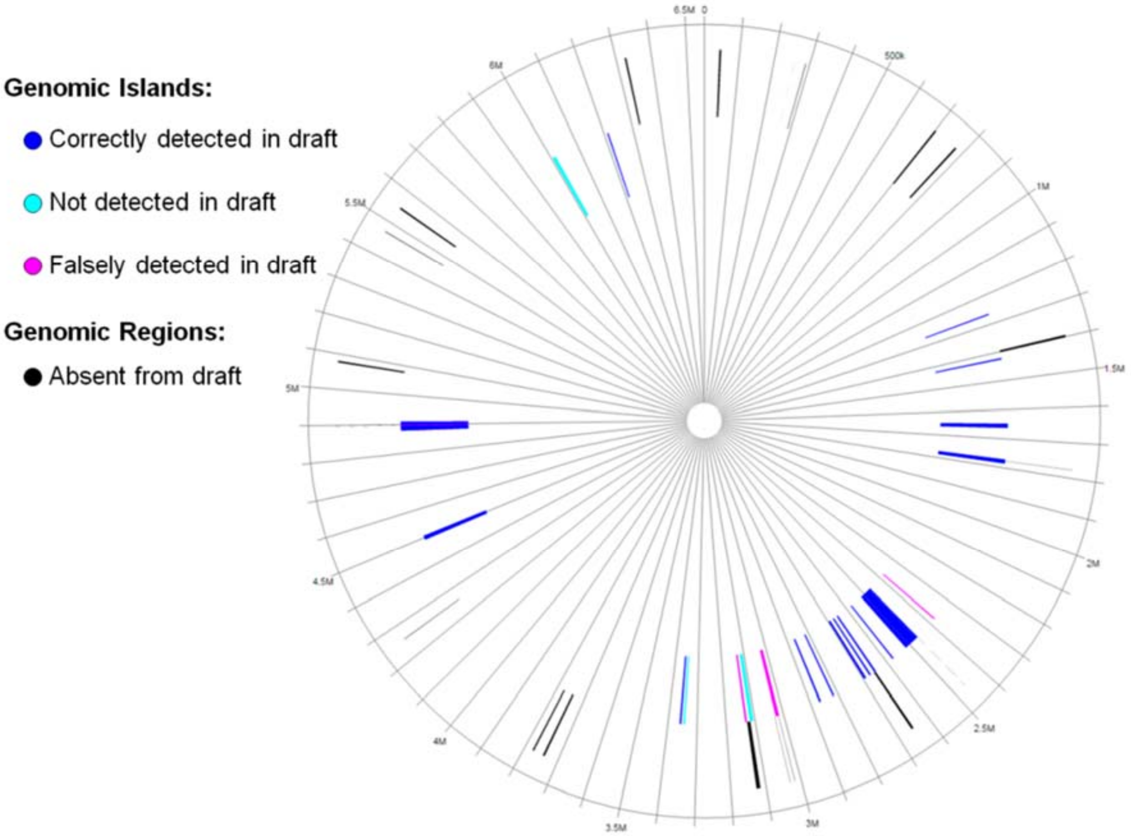


Figure 2.15 Example: Simpler visualization of GI predictions in *Pseudomonas aeruginosa* LES431. This visualization shows that the GI at 4,878,834 bp was correctly predicted.

2.5.3. *Listeria monocytogenes* results

Results for the comparison of GI prediction in draft and complete GI predictions are shown in Table 2.3. Note that IslandPath-DIMOB results are from a new version that will be released with a new version of IslandViewer, rather than the version of IslandPath-DIMOB that is included in the current version of IslandViewer described in the next chapter. Sensitivity and precision for each GI prediction method, and for both of the measurements described in 2.5.1, are shown. Although accuracy can be calculated with the per base measurement of TNs, it was found that accuracy was always close to 1 because of the high number of TNs, and was not informative. Both methods have very high precision, and SIGI-HMM has high sensitivity. While IslandPath-DIMOB has a lower sensitivity, it is still high enough to be in line with other GI prediction methods for complete genomes (Langille et al., 2008a). Sensitivity and precision of IslandViewer predictions were compared with N50 for each of the *L. monocytogenes* genomes, as shown in figures 2.16 and 2.17. No clear link between N50 and sensitivity or precision was found for this data set, but this may have been because all of the genomes were of relatively high quality, and each genome had a high N50.

Table 2.3 GI prediction sensitivity and precision for draft *Listeria monocytogenes* genomes

		Sensitivity	Precision
SIGI-HMM	50% overlap	0.9867	1.0000
	per base	0.9292	0.9712
IslandPath-DIMOB	50% overlap	0.7627	0.9864
	per base	0.8210	0.9585

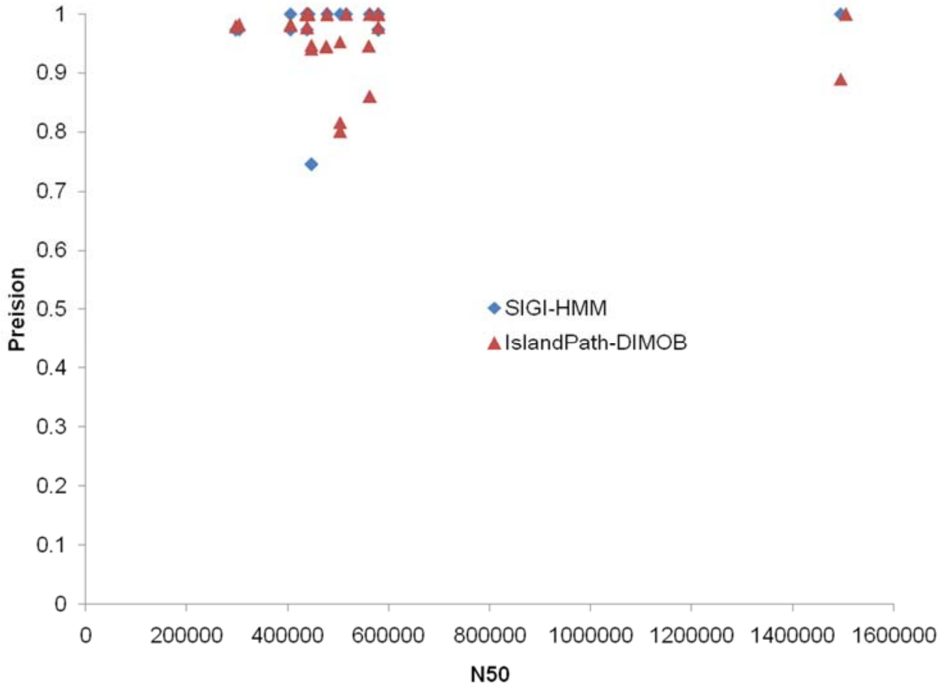


Figure 2.16 **Graph of N50 vs precision for GI predictions in the *Listeria monocytogenes* data set. Each point represents predictions from one method for one genome. No correlation was observed.**

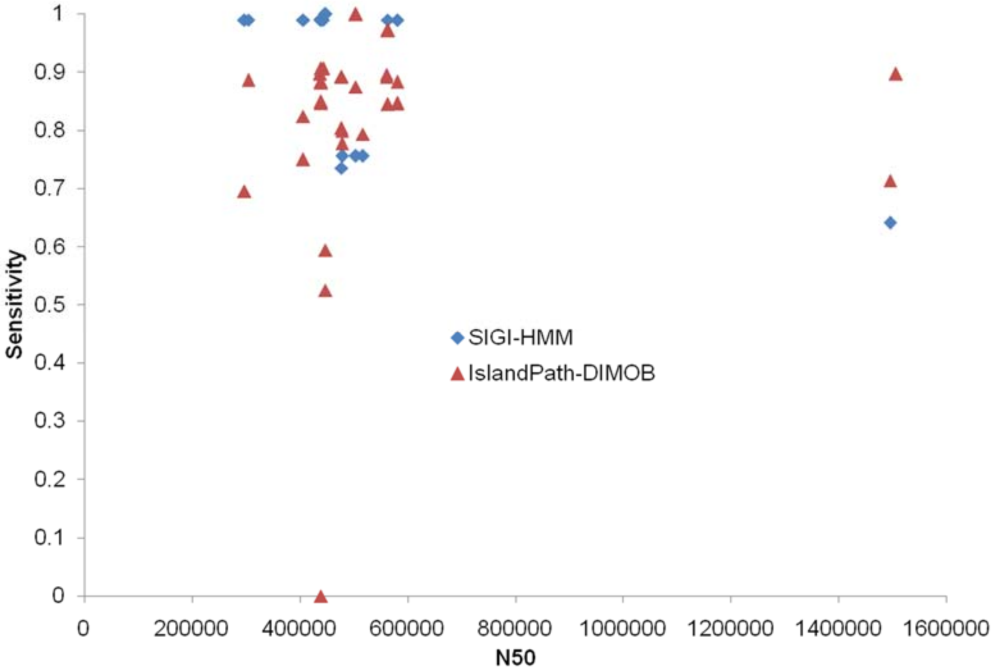


Figure 2.17 **Graph of N50 vs sensitivity for GI predictions in the *Listeria monocytogenes* data set. Each point represents predictions from one method for one genome. No correlation was observed.**

To determine whether there is a relationship between IslandViewer performance and gaps in draft genomes, IslandViewer predictions in the complete *L. monocytogenes* genomes were binned according to how close they were to the edge of a contig in the draft genome, and this grouping is shown in Figure 2.18. This was done with combined SIGI-HMM and IslandPath-DIMOB predictions for the 36 complete genomes, where overlapping predictions were combined to form a single GI prediction. If a GI prediction overlapped with the edge of a contig, including overlapping with regions not covered in the draft genome, the distance from a contig edge was considered to be 0. GIs were also separated according to whether they were correctly predicted in the draft genome, based on the 50% overlap method described in section 2.5.1. This separation of GIs correctly predicted or missed in draft genome analysis is also shown in Figure 2.18. While not all GI predictions occur where there are gaps in the draft version of a genome, and not all GI predictions that occur at these gaps are missed, a non-negligible proportion of GIs do occur at these gaps, and these GIs were missed more often in draft genome analysis than those which are further away from gaps for this data set.

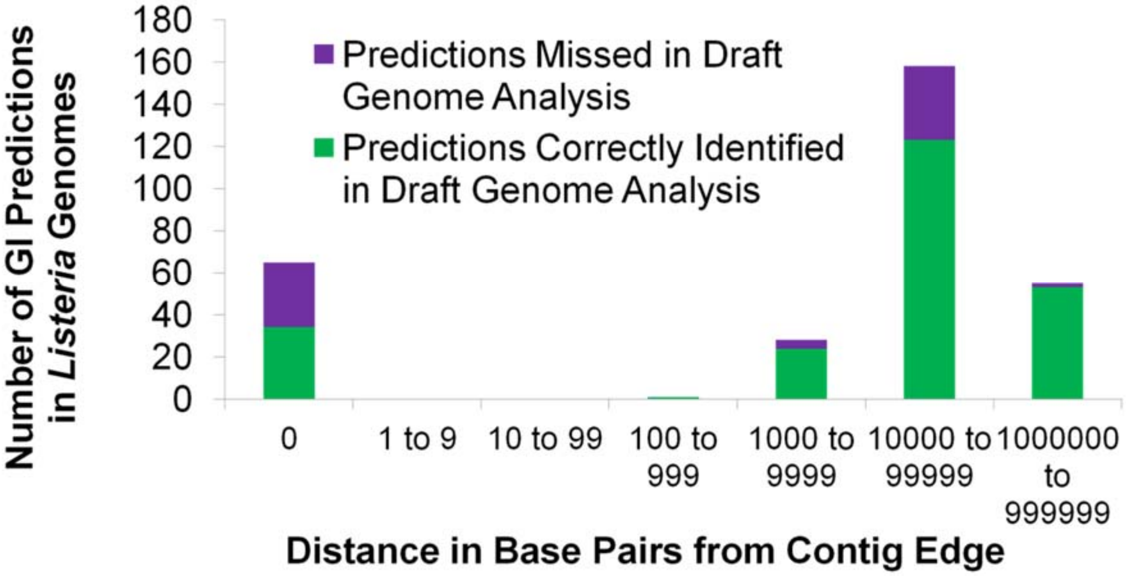
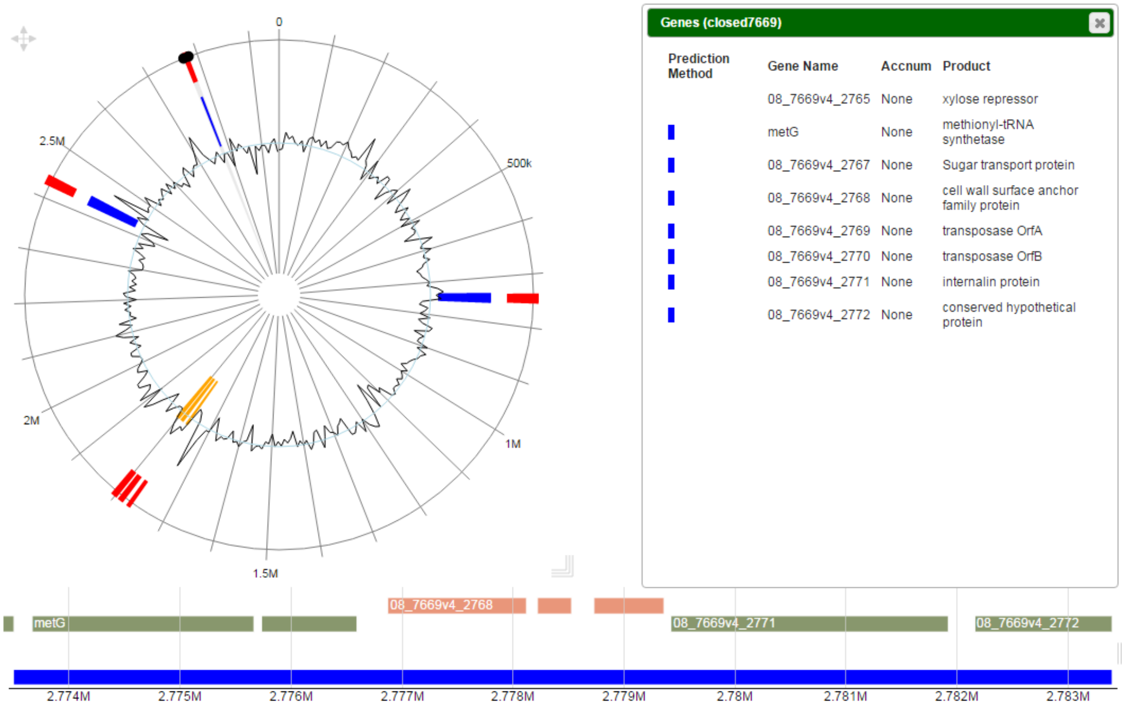


Figure 2.18 Integrated GI predictions in complete *Listeria monocytogenes* genomes, binned by distance from a contig edge in the draft genome, and by whether the GI was correctly predicted in the draft genome.

Figure 2.19 shows the IslandViewer interface, which will be described in detail in section 3.4.2, zoomed in on a GI which was detected in the complete *L. monocytogenes* 08-7669 genome, but missed in the draft version of the genome. The circular view shows the whole genome (with the zoomed in region denoted by black dots), and the zoomed in region is shown in the linear view. Note that because the new version of IslandPath-DIMOB is not yet integrated into the IslandViewer interface, this image shows IslandPath-DIMOB predictions made using IslandViewer 3. The highlighted GI is also predicted in the most up-to-date version of IslandPath-DIMOB. Two contig edges are contained within this GI prediction: one at 2778324 bp and the other at 2779323 bp (not shown). This likely contributed to the GI being missed in the draft genome. This example is also notable because the contig edges are occurring at the locations of two transposases within the GI.



Prediction Method	Gene Name	Accnum	Product
	08_7669v4_2765	None	xylose repressor
▮	metG	None	methionyl-tRNA synthetase
▮	08_7669v4_2767	None	Sugar transport protein
▮	08_7669v4_2768	None	cell wall surface anchor family protein
▮	08_7669v4_2769	None	transposase OrfA
▮	08_7669v4_2770	None	transposase OrfB
▮	08_7669v4_2771	None	internalin protein
▮	08_7669v4_2772	None	conserved hypothetical protein

Figure 2.19 An example of a GI that was missed during draft genome analysis in the IslandViewer interface. The blue bands signify IslandPath-DIMOB predictions. Two contig edges and two transposase genes are contained within the GI.

2.5.4. *Pseudomonas aeruginosa* results

Results for the comparison of GI prediction in draft and complete GI predictions are shown in Table 2.4. Sensitivity and precision for each GI prediction method, and for both of the measurements described in 2.5.1, are shown. Sensitivity and precision for each method is lower than for the *L. monocytogenes* data set. This is might be because most draft genomes are not being compared with a complete genome from the same isolate, and the accessory genomic structure of *P. aeruginosa* is highly diverse (Klockgether et al., 2011). Differences between draft and complete genomes being compared here may be due to true differences between isolates, as opposed to errors or missing regions in the draft genome.

Table 2.4 GI prediction sensitivity and precision for draft *Pseudomonas aeruginosa* genomes

		Sensitivity	Precision
SIGI-HMM	50% overlap	0.5543	0.7588
	per base	0.5346	0.6979
IslandPath-DIMOB	50% overlap	0.5959	0.8275
	per base	0.5919	0.8158

IslandViewer predictions from complete *P. aeruginosa* genomes were binned according to how close they were to the edge of a contig in the draft genome, similarly to the binning of *L. monocytogenes* GI predictions described in the previous section, and this binning is shown in Figure 2.20. This figure reflects the low overall sensitivity of IslandViewer for the *Pseudomonas* data set, but it also shows that, like in the *Listeria* data set, many GI predictions overlap with contig edges in this data set. Note that, as mentioned in previous sections, GI predictions from some complete *Pseudomonas* genomes are counted multiple times because they are being compared to more than one draft genome.

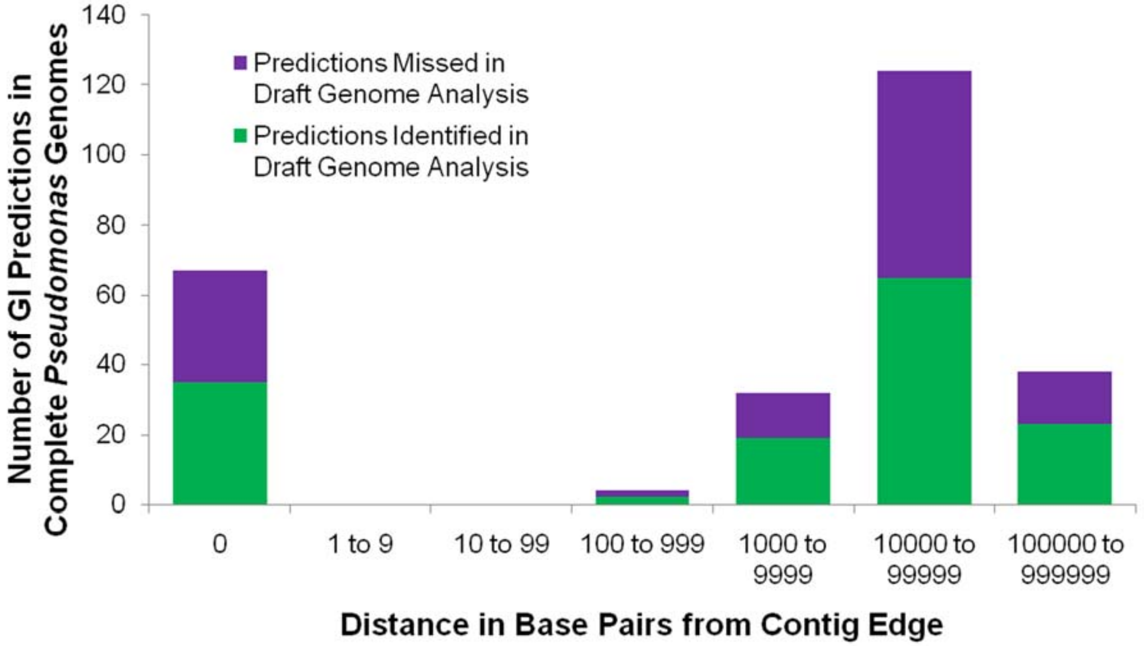


Figure 2.20 Integrated GI predictions in complete *Pseudomonas aeruginosa* genomes, sorted by distance from a contig edge in the draft genome, and by whether the GI was correctly predicted in the draft genome.

2.5.5. Antisense transcription

An additional analysis of GIs was performed on *Citrobacter rodentium* and *Helicobacter pylori.* Directional RNA-Seq data was used to perform a genome-wide assessment of cis-antisense transcription in GIs. Normal sequencing preparation protocols include an amplification step that leads to a loss of strand-specific information in RNA-Seq libraries. There are multiple methods for directional RNA-Seq which use sequencing preparation protocols that maintain strand specificity (Mamanova et al., 2010; Ozsolak et al., 2009; Wu et al., 2008). The data sets were downloaded from the NCBI SRA database (the SRA Study IDs are ERP000493 for the *Citrobacter rodentium* data set and SRP001481 for the *Helicobacter pylori* data set). At the time of the analysis, there were very few directional RNA-Seq data sets, so the species examined were chosen based on the availability of data.

dRNA-Seq reads were aligned to a reference genome with Bowtie (Langmead et al., 2009), GIs were predicted using IslandViewer (Dhillon et al., 2013), and ORFs were predicted in the reference genome using Glimmer (Delcher et al., 1999). A custom script was used to determine the percentage of ORFs with above average antisense transcription inside GIs vs. outside of GIs. There was no increase in antisense transcription in predicted GIs for either of the data sets examined. This result does not support the hypothesis that antisense transcription occurs disproportionately in GI regions.

Table 2.5 Percentage of ORFs with antisense transcription in each data set, both inside and outside of GIs

	GIs	non-GI regions
Citrobacter rodentium	0.1938	0.8061
Helicobacter pylori	0.0123	0.0137

2.6. *Listeria* vs. *Pseudomonas* analysis: common characteristics

Both the *Listeria* and *Pseudomonas* data sets consisted of high quality draft genomes, with high N50 values and a high percentage of total CDSes included in the draft. All of the *Listeria* genomes, and most of the *Pseudomonas* genomes, were sequenced using Illumina short read sequencing technology. The two sets of draft genomes also have similar biases in terms of what they were missing. Both data sets found that VFs and AMR genes are not overrepresented in missing regions. In the *Pseudomonas* data set, VFs were even found to be underrepresented in these missing regions. These results are promising for the use of WGS for clinical and epidemiological purposes. That being said, any researcher, clinician, or epidemiologist that works with draft genome data should always be aware that it is possible for a gene of interest to be missing from a draft genome.

In both data sets, tRNA genes and CDSes assigned to the replication, recombination, and repair superfamily are significantly overrepresented in regions missing from draft genomes. CDSes which could not be assigned to a COG category are also highly overrepresented in missing regions of draft genomes in both data sets. No other COG superfamily was overrepresented in either data set, although as discussed in 2.4.4, this may be due the extremely high proportion of unassigned CDSes. These overrepresented groups are of particular interest because they are all associated with GIs (Hsiao et al., 2005).

It is unsurprising that tRNA genes are commonly missing from draft genomes; tRNA genes are repetitive, which makes them more difficult to sequence. The generalized structure of tRNA can be described as a cloverleaf, with four loops and at least four stems held together by Watson-Crick base pairs (Holley et al., 1965). Therefore, tRNA genes contain multiple pairs of sequences which are reverse complementary to each other, with one pair for each stem. This pattern of reverse complementary sequences can be used to computationally predict tRNA genes in DNA sequences (Marvel, 1986). tRNA genes also have high sequence similarity to each other, and can cluster together (Vold, 1985). Repetitive genomic regions are more difficult to sequence (see section 1.1.3). Repetitiveness both within and between tRNA genes may both be contributing to the low coverage of tRNA genes in draft genomes.

GIs were able to be predicted well in the *L. monocytogenes* draft genomes, but less well in the *P. aeruginosa* data set. However, analysis of the *Pseudomonas* data set was limited by not having complete genomes from identical isolates for comparison. In both data sets, a large number of GI predictions overlap with contig edges or regions that are missing from draft genomes, and in the *Listeria* data set these GIs were less likely to be correctly predicted in draft genomes. When performing comparative genomics analysis of GI regions, it is useful to note that GIs which are present in an isolate's genome may be hidden by a sequence gap.

Results of this GI analysis may be biased due to the GI prediction tools in IslandViewer. In particular, IslandPath-DIMOB only reports GIs which contain a mobility gene such as a transposase, and transposases are disproportionately missing from draft genomes. Different GI prediction tools focus on different compositional biases (see section 1.3.1) and could identify different GIs with characteristics that make them more or less likely to be missing from draft genomes. Also, the current version of IslandPath-DIMOB does not incorporate a method to optimize prediction boundaries, and this may have an impact on the results of this analysis. Differing boundaries of GI predictions that are otherwise in agreement between draft and complete genomes have an impact on the "per base" method of calculating TPs, FPs, and FNs described in 2.5.1

The combined results that some GIs overlap with contig boundaries and missing regions, and that tRNA genes and transposases are disproportionately missing from draft genomes suggest that complete genomes are ideal, particularly for comparative genomics studies of GIs in these species. That being said, IslandViewer still overall performs well when using draft genomes as input, at least in the *Listeria* data set, for which complete genomes from identical isolates are available for comparison. The differences between the *Listeria* and *Pseudomonas* data sets highlight the need for this analysis to be repeated on other species of particular interest in the future to ensure that the utility of draft genomes for functional analyses is well understood.

Chapter 3.

Improving IslandViewer including the ability to process draft genomes

Portions of this chapter have been previously published in the following article 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. I., & Brinkman, F. S. L. (2015). IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Research*, *43*(W1), W104–8. http://doi.org/10.1093/nar/gkv401

Note: IslandViewer 3 development was led by Bhavjinder Dhillon and Matthew Laird. I developed the draft genome analysis pipeline, performed the assessment of IslandViewer 3 prediction methods, and aided the IslandViewer 3 interface development.

3.1. Background and rationale

This chapter describes IslandViewer 3, a major new release of IslandViewer software with several updates. IslandViewer is a web-based genomic island (GI) prediction tool which incorporates three of the most accurate GI prediction methods: IslandPick, IslandPath-DIMOB, and SIGI-HMM (Hsiao et al., 2003; Langille et al., 2008a; Waack et al., 2006). For more information about IslandViewer, see section 1.3.4. At the start of this project, there was a lack of tools for GI prediction in draft genomes. One tool did exist, but the website for this tool was shut down shortly after the research article describing the tool was published (Lee et al., 2013). However, there was high demand for a GI prediction tool for draft genomes. As described in section 1.1.5, the vast majority of bacterial genomes are only being sequenced to the draft stage. Prior to the release of IslandViewer 3, one of the most requested features was the ability to analyze draft genomes. To accommodate this, IslandViewer now accepts draft genomes as input for its analysis pipeline.

The other new features in IslandViewer 3 are an upgraded backend, a new visualization tool, and updated gene annotations. Upgrades to the IslandViewer backend and visualizations in IslandViewer 3 were prompted by the increased number of complete genomes for which IslandViewer stored pre-computed results, and the increased number of genomes being uploaded for custom analysis. Due to these updates, IslandViewer requires less data storage per genome and will not crash due to a high volume of requests for custom analysis. As well as reducing data storage, the dynamic visualization of IslandViewer 3 also improves the user experience. Annotations of AMR genes, VFs, and pathogen-associated genes were updated and expanded to increase the functionality of IslandViewer for the analysis of clinical and epidemiological isolates.

The main focus of this chapter is the ability of IslandViewer to accept draft genomes as input. Other updates in IslandViewer 3, including an updated backend, a new visualization tool, and updated annotation, are also described.

3.2. Draft genome pipeline

3.2.1. Pipeline overview

The IslandViewer draft genome pipeline requires an assembled, annotated draft genome sequence and a user-selected reference genome as input. The uploaded contigs are aligned against the user-selected reference genome using MCM (Rissman et al., 2009), and then a single concatenated sequence is generated based on this alignment. In cases where the reverse complement of a contig is aligned to the reference sequence, the reverse complement of that contig is included in the concatenated sequence instead of the original contig sequence. Any unaligned contigs are included at the end of the concatenated sequence for analysis, but they are clearly labeled as unaligned contigs in the IslandViewer output. This ‘concatenated-contigs genome’ is then run through the existing IslandViewer pipeline. IslandPick is not run on draft genomes due to the possibility of incorrectly oriented contigs (see section 1.3.2 for more information about the IslandPick approach). This simple approach allows for the identification of GI predictions in draft genomes, with all its caveats.

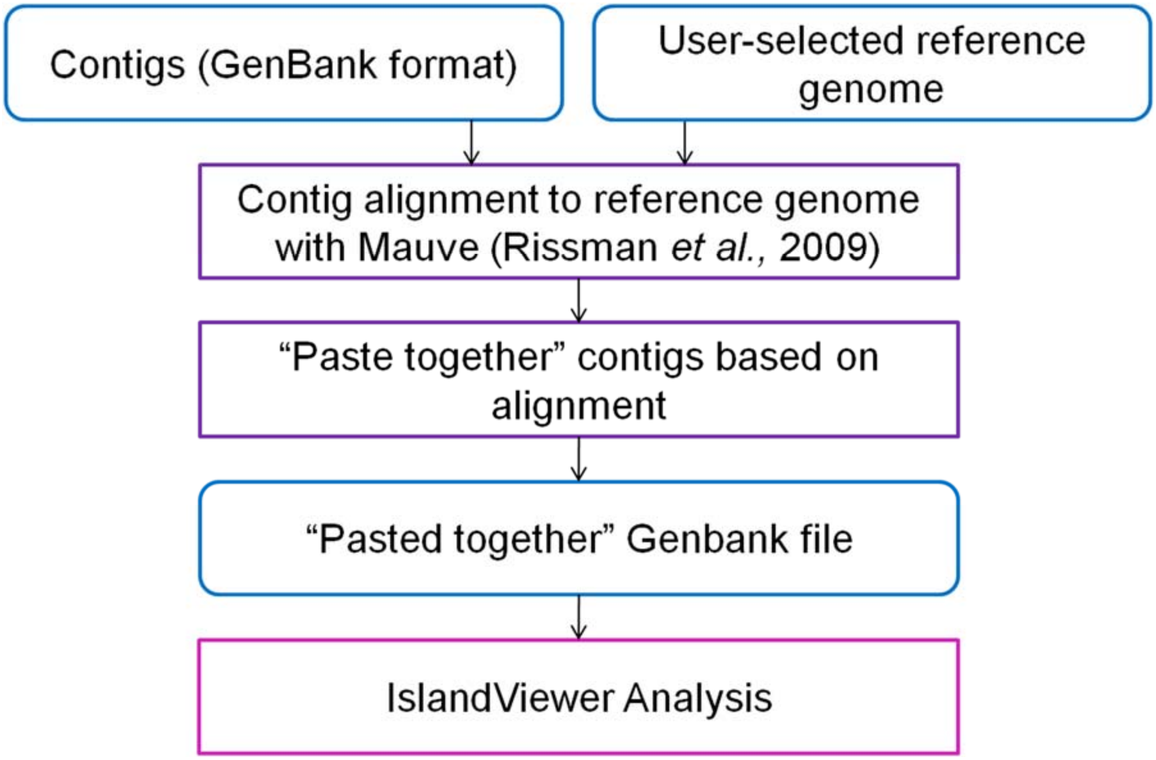

Contigs (GenBank format)
User-selected reference genome
Contig alignment to reference genome with Mauve (Rissman *et al.*, 2009)
“Paste together” contigs based on alignment
“Pasted together” Genbank file
IslandViewer Analysis

Figure 3.1 Flowchart describing IslandViewer draft genome pipeline

3.2.2. Contig ordering

IslandViewer uses the Mauve Contig Mover (MCM) to arrange contigs against a reference genome sequence in the draft genome analysis pipeline. MCM was built on the framework of Mauve, a multiple genome alignment tool (Darling et al., 2004). Mauve uses locally collinear blocks to align multiple genomes. These locally collinear blocks are highly similar local sequence alignments which act as anchors around which the rest of the sequence is aligned. MCM performs iterative alignments between the set of concatenated contigs and the reference genome. After each alignment stage, MCM rearranges contigs in order to maximize the lengths of locally collinear blocks. That is, it detects contig edges that correspond to locally collinear block edges, and rearranges those contigs such that their edges are put together. Contigs may also be converted to their reverse complement sequence in order to maximize locally collinear blocks. MCM then repeats the alignment and rearrangement steps until no further rearrangements can further optimize the alignment. With each alignment stage, MCM produces several output files. The output files do not include a single sequence of the concatenated set of contigs, but it does include two files that are used in the concatenation step of the IslandViewer draft genome pipeline. The first of these files is a list of contigs with information about their order, location, and orientation in the alignment, and the second is an alignment backbone file which lists exactly which coordinates in the draft genome aligned to which coordinates in the complete reference. The backbone file is used to separate contigs which are aligned to the complete reference from those which are not successfully aligned.

3.2.3. Concatenation

After MCM determines the optimum arrangement of contigs, a custom Perl script is used to produce a concatenated genome sequence in GenBank format. This concatenated genome file appears as a single, complete genome sequence to be used as input for IslandPath-DIMOB and SIGI-HMM. The locations of individual contigs within the concatenated sequence are listed in the annotations of this GenBank file, and this file can be downloaded through the IslandViewer interface. Coordinate information for all annotations of the original contigs are adjusted accordingly and incorporated into the concatenated Genbank file. Contigs which were not aligned to the reference sequence in the MCM step are appended to the end of the concatenated sequence. Contigs are labelled by their alignment status in the IslandViewer interface (see section 3.5.3 for more details about the IslandViewer interface for draft genomes).

Within the concatenated sequence, contigs are separated by a constant 1000 bps of unassigned bases ("n"). An earlier version of the draft genome pipeline contained an extra step which calculated specific lengths of unassigned base sequences to separate contigs. In this earlier version, the length of sequence separating contigs was set to the length of the complete reference sequence which separated the regions that aligned to the two adjacent contigs, with a minimum of 3 bp and a maximum of 10 kbp of separating sequence. However, the length of separating sequence had no impact on IslandViewer predictions in the *Listeria monocytogenes* draft genome data set described in 2.3.1. The switch to a constant length of separating sequence saves an extra computational step, and ensures that contig boundaries are easily visible in the IslandViewer interface. Other researchers who have produced concatenated contig sequences used a sequence which contains stop codons in all six reading frames. This was used to prevent ORFs which span contigs from being detected by annotation tools (Athey et al., 2016), and was determined to be unnecessary for the purposes of IslandViewer, because contigs are annotated prior to concatenation.

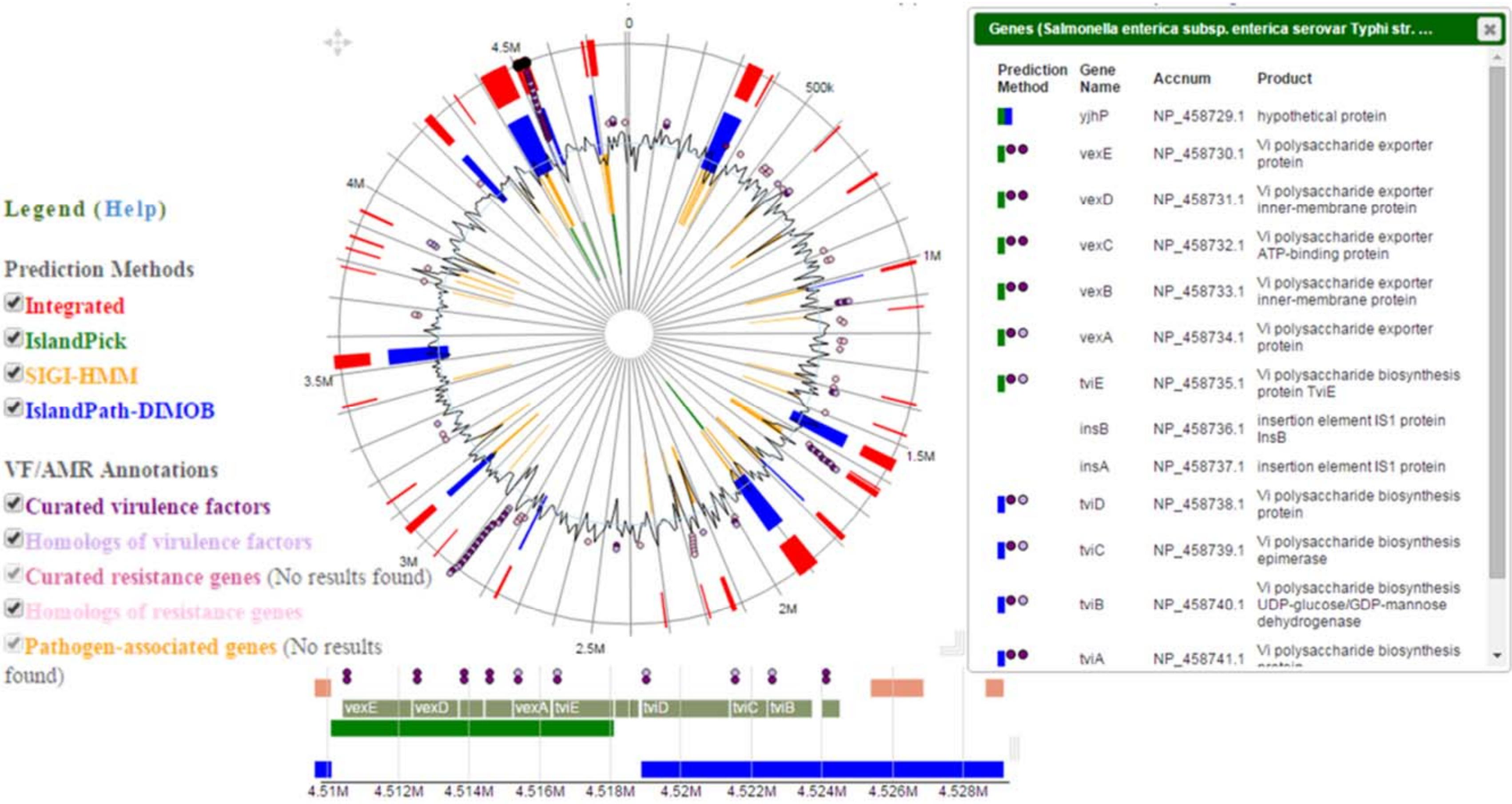
3.3. Other new features

3.3.1. Backend upgrades

Earlier versions of IslandViewer were built upon MicrobeDB (Langille et al., 2012), a database of all complete bacterial genomes as downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). As the number of genomes continued to grow, the coupled IslandViewer and MicrobeDB database became very inefficient. The new version of IslandViewer is completely separated from MicrobeDB and only stores information relevant to IslandViewer users. New backend update scripts reflect changes in file structure and genome annotations in NCBI, and allow for monthly updates to pre-computed analyses so that they include new complete bacterial or archaeal genomes. In addition to this, IslandViewer can now submit custom analyses with a robust queuing system. This improves processing time of custom analyses of existing genomes (for example, using different IslandPick comparative genomics criteria) and analysis of new user-uploaded genomes, and virtually eliminates the most common errors prevalent in the former system.

3.3.2. Visualization with GenomeD3Plot

GenomeD3Plot, which was used to create the visualizations described in section 2.5.2, was developed for IslandViewer 3 as a new genome visualization library based on the D3 javascript library (http://www.d3js.org) (Laird et al., 2015). With this new tool, interactive circular, horizontal and vertical genome views are provided in IslandViewer for both user-uploaded and pre-computed genome analyses. Importantly, GenomeD3Plot is able to generate visualizations dynamically, eliminating the need to store pre-computed images for every permutation of information to display, as was done previously. This greatly improves IslandViewer's ability to handle the increasing number of bacterial and archaeal genomes while minimizing storage requirements and provides a richer, more interactive genome browsing experience. Within the three separate views, GI predictions are shown, broken down by prediction method, along with annotations of virulence factors, antimicrobial resistance genes and pathogen-associated genes. Users can specify which GI prediction method and annotations to display, and can select regions in the circular view to zoom in or out, which updates the horizontal and vertical genome on the selected region. While the horizontal viewer provides a more detailed visual representation of a selected region of the genome, the vertical viewer provides text descriptions of the genes and gene products located within the genomic region of interest. Both horizontal and vertical viewers have their own zoom/navigation features (using a mouse scroll wheel, for example over the horizontal view). For pre-computed genome analyses, both viewers provide links to the NCBI for any selected gene, or in the case of virulence and antimicrobial resistance gene annotations, links are provided to the source of the annotation for more information. A side-by-side comparison of two genomes is also supported by GenomeD3Plot. Additionally, as users navigate through a genome of interest, a page can be ‘saved’ for linking back to selected regions and zoom levels using a unique URL.



Legend (Help)

Prediction Methods

☑ Integrated
☑ IslandPick
☑ SIGI-HMM
☑ IslandPath-DIMOB

VF/AMR Annotations

☑ Curated virulence factors
☑ Homologs of virulence factors
☑ Curated resistance genes (No results found)
☑ Homologs of resistance genes
☑ Pathogen-associated genes (No results found)

Genes (Salmonella enterica subsp. enterica serovar Typhi str. ...

Prediction Method	Gene Name	Accnum	Product
🟩🟦	yjhP	NP_458729.1	hypothetical protein
🟩●●	vexE	NP_458730.1	Vi polysaccharide exporter protein
🟩●●	vexD	NP_458731.1	Vi polysaccharide exporter inner-membrane protein
🟩●●	vexC	NP_458732.1	Vi polysaccharide exporter ATP-binding protein
🟩●●	vexB	NP_458733.1	Vi polysaccharide exporter inner-membrane protein
🟩●○	vexA	NP_458734.1	Vi polysaccharide exporter protein
🟩●○	tviE	NP_458735.1	Vi polysaccharide biosynthesis protein TviE
	insB	NP_458736.1	insertion element IS1 protein InsB
	insA	NP_458737.1	insertion element IS1 protein
🟦●○	tviD	NP_458738.1	Vi polysaccharide biosynthesis protein
🟦●○	tviC	NP_458739.1	Vi polysaccharide biosynthesis epimerase
🟦●○	tviB	NP_458740.1	Vi polysaccharide biosynthesis UDP-glucose/GDP-mannose dehydrogenase
🟦●●	tviA	NP_458741.1	Vi polysaccharide biosynthesis

Figure 3.2 The new IslandViewer visualization with GenomeD3Plot, in this example it is showing the chromosome of *Salmonella enterica* subsp. enterica serovar Typhi str. CT18

Because of the interactive nature of GenomeD3Plot, IslandViewer 3 also allows users to search for particular genes of interest within a genome, and will highlight the genes of interest in each view. This function allows users to better navigate through a genome of interest. Figure 3.2 shows the GenomeD3Plot visualizations for Salmonella *enterica* subsp. enterica serovar Typhi str. CT18 after searching for a known virulence factor, *vexE*. This gene is focused in all views and highlighted in the vertical panel to easily evaluate the GI predictions and annotations of *vexE* and its neighboring genes. This GenomeD3Plot viewer and associated gene search functionality can serve as a broader search tool to study a genome of interest at various levels of detail, including virulence, resistance or pathogen-associated genes.

3.3.3. Updated annotations

In the previous IslandViewer update, 956 antimicrobial resistance gene annotations were incorporated from the ARDB (Liu & Pop, 2009). To further improve resistance gene annotations, all proteins in IslandViewer 3 precomputed results as of November 2014 were analyzed using the most precise version of the previously published RGI method for identifying genes involved in AMR (McArthur et al., 2013). The AMR gene annotations from ARDB have been curated and incorporated into CARD, thus, these new annotations have replaced the previous ARDB annotations in IslandViewer. Exact matches to curated antimicrobial resistance genes are denoted as curated resistance genes and colored in pink in IslandViewer 3. Any hits that were found within the strict criterion are denoted as homologs and are displayed in light pink. Through this analysis, antimicrobial resistance gene annotations for pre-computed genomes have been greatly expanded from 956 to 28 911 resistance genes, providing an overview of predicted molecular antimicrobial resistance profiles for 589 distinct genera available to date in IslandViewer.

The previously published pathogen-associated genes analysis was updated using the same methodology as outlined by Ho Sui et al. in 2009 (Sui et al., 2009). The original analysis was performed on 631 genomes, of which 298 were pathogens. An update of this analysis to include every genome completely sequenced up to September 2014 (and available through NCBI) required manual curation of 2794 genomes as either from a pathogen or non-pathogen, using the same criteria accepted in the previous analysis. In total, 1277 pathogen and 1517 non-pathogen genomes were compared, to determine the set of genes currently specific for pathogen genomes using set criteria, termed pathogen-associated genes. These pathogen-associated genes have been shown to be disproportionately involved in more ‘offensive’ virulence roles such as invasion into a host, type III/IV secretion systems, or toxins, rather than defence roles. Such genes are of interest since they may represent novel virulence factors under certain conditions. After this updated analysis, 18 919 pathogen-associated genes (found in three or more distinct genera) were identified and annotated in IslandViewer 3. All results from this analysis are also available for reference at http://pathogenomics.sfu.ca/pathogen-associated/2014.

To aid identification of pathogenicity islands within GI predictions, an expanded virulence factor annotation was complete. More than 1600 curated virulence factors were annotated in the last release of IslandViewer using the VFDB (Chen et al., 2012). Since then, over 8000 additional curated virulence factor gene annotations have been collected from the expanded Virulence Factor Database (VFDB), PATRIC (Wattam et al., 2014) and Victor's virulence factors (http://www.phidias.us/victors/) and mapped to their corresponding proteins in IslandViewer 3. Only a subset of virulence factor annotations from PATRIC, those with curated links to literature, were incorporated. These annotations are displayed in purple in the genome visualizations.

However, such curated virulence factors still only cover a limited number of genomes and as the number of very closely related genomes sequenced increases, it is clear that many of these curated virulence factors should also be annotated in highly similar genomes of the same species or serovar. To address this issue, a very conservative reciprocal best blast hit (RBBH) approach was used to identify homologs (essentially probable orthologs) of curated virulence factor genes using strict criteria: the virulence factor annotation transfer was only permitted if the gene occurred within the same species, plus the CVTree (Xu & Hao, 2009) distance between the genomes being compared was <0.3 (to ensure that annotation transfer did not occur even within a species if the genomes were more divergent). Additional filters were placed specifically upon genera with notable phenotypic variability within the species, e.g. *Salmonella* and *Escherichia* genera, such that annotation transfer was only permitted between genomes from the same serovar or strain for such species. Annotations were also only transferred if the RBBH BLASTp e-value was lower than 1e–10, plus the sequences shared ≥90% sequence identity, plus the BLAST hit (high scoring segment pair) also covered at least 80% of the query sequence length. These very stringent criteria were selected in order to maximize precision/specificity for the annotation of virulence factor gene homologs at the expense of recall/sensitivity to ensure annotations would be most likely correct, at the expense of missing some. Even though this criteria tends to identify orthologs by widely accepted RBBH criteria, they are referred to as homologs in IslandViewer 3 and annotated with a different lighter purple color in the genome visualization to highlight that they are not confirmed, curated virulence factors. Using this approach, an additional 39 441 virulence gene homologs in 485 genomes, covering 37 distinct pathogen genera were annotated. With this expanded dataset, users can view and explore the presence/absence of PAIs in many more genomes than previously, including very closely related genomes from different strains of a species which clearly contain the same classic virulence factors for that species. Of course, in the end such annotations should always be thought of as an initial guide, or hypothesis-generating for more in depth future analysis, due to the highly contextual nature of virulence.

3.4. Implementation into IslandViewer

3.4.1. Evaluation of GI prediction methods

During the beta testing phase of the new version of IslandViewer, GI predictions from IslandPath-DIMOB and IslandPick were found to be different from those predicted using the previous IslandViewer version. These methods were not meant to have been changed from the previous version of IslandViewer. IslandPick selects comparison genomes to use for GI detection (Langille et al., 2008a), and predicted GIs may differ depending on the comparison genomes chosen. Differences in available genomes for comparison were a possible cause for the different results observed. However, this cause was ruled out by redoing analyses on the previous version of IslandViewer with an updated set of genomes available for comparison. The changes in each method were identified, and an assessment of each method's performance with and without these changes was performed. The methodology of a previous assessment of GI prediction tools, including IslandPath-DIMOB and IslandPick, was used again for this assessment (Langille et al., 2008a).

Differences in IslandPath-DIMOB results were due to a switch from HMMER2 to HMMER3 for identification of mobility genes. HMMER3 is much faster than HMMER2, and is more sensitive but has a slight reduction in specificity (Eddy, 2011). Both versions of IslandPath-DIMOB were assessed using the same set of 117 genomes used in the previous assessment (Langille et al., 2008a). Precision, sensitivity, and accuracy were calculated using the equations and per-base method described in section 2.5.1. The set of positives were the original IslandPick predictions in these genomes, and the set of negatives were highly conserved regions. Table 3.1 shows the performance of IslandPath-DIMOB with HMMER2 and with HMMER3. Accuracy was similar between the two versions, but precision was slightly lower and sensitivity slightly higher when using HMMER3. This is unsurprising, as this result mirrors the reported differences between HMMER2 and HMMER3. Due to its significantly increased speed, HMMER3 continues to be used for IslandPath-DIMOB prediction in IslandViewer 3.

Table 3.1 IslandPath-DIMOB performance using HMMER2 and HMMER3

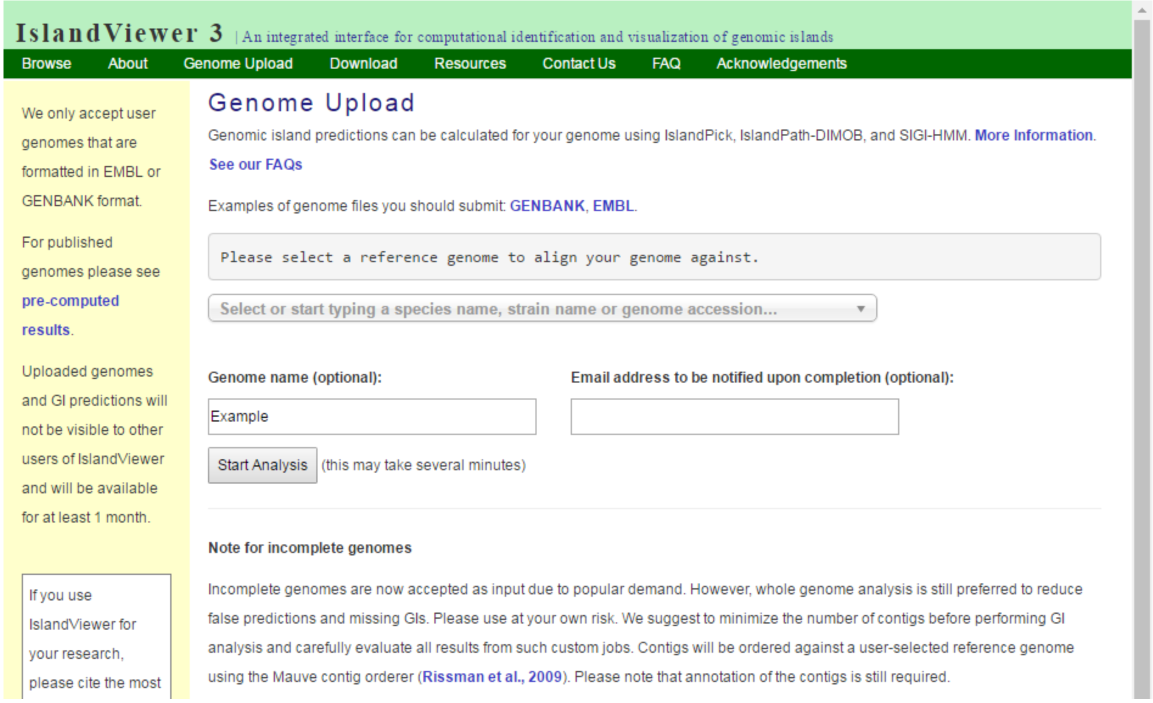
	Precision	Sensitivity	Accuracy
Results from Original IslandViewer (HMMER2)	0.865	0.355	0.862
Results from IslandViewer 3 (HMMER3)	0.810	0.386	0.861

The cause for differences in IslandPick results was identified as a minor change in the comparison genome picking algorithm in IslandViewer 2 which was inadvertently corrected in IslandViewer 3. The minimum single close genome distance cut off requires that at least one genome used for IslandPick comparison has a CVtree distance of at least that amount from the query genome (Langille et al., 2008a; Xu & Hao, 2009). That minimum distance was reported as 0.34, but was set to 0 in IslandViewer 2. To determine which minimum distance currently produces the best results, I repeated an assessment using literature-derived GIs from 5 genomes. The results from this assessment are shown in Table 3.2. While IslandPick had perfect precision with either distance value, both sensitivity and accuracy were higher when using a minimum distance of 0.34. This higher minimum distance was implemented in the release of IslandViewer 3. Note that over time, as more genomes become available for comparison, the IslandPick parameters for choosing comparison genomes should be adjusted to ensure that the best comparison genomes are chosen.

Table 3.2 IslandPick performance using different minimum single close genome distances

	Precision	Sensitivity	Accuracy
Results from original IslandViewer (Langille et al., 2008a)	1.000	0.870	0.963
IslandViewer 3 Minimum Distance of 0	1.000	0.725	0.924
IslandViewer 3 Minimum Distance of 0.34	1.000	0.798	0.962

3.4.2. Draft genome submission


IslandViewer 3 | An integrated interface for computational identification and visualization of genomic islands
Browse
About
Genome Upload
Download
Resources
Contact Us
FAQ
Acknowledgements

We only accept user genomes that are formatted in EMBL or GENBANK format.

For published genomes please see **pre-computed results**.

Uploaded genomes and GI predictions will not be visible to other users of IslandViewer and will be available for at least 1 month.

If you use IslandViewer for your research, please cite the most

Genome Upload

Genomic island predictions can be calculated for your genome using IslandPick, IslandPath-DIMOB, and SIGI-HMM. **More Information**.

See our FAQs

Examples of genome files you should submit: **GENBANK**, **EMBL**.

```
Please select a reference genome to align your genome against.
```

Select or start typing a species name, strain name or genome accession... ▾

Genome name (optional): Example

Email address to be notified upon completion (optional):

Start Analysis (this may take several minutes)

Note for incomplete genomes

Incomplete genomes are now accepted as input due to popular demand. However, whole genome analysis is still preferred to reduce false predictions and missing GIs. Please use at your own risk. We suggest to minimize the number of contigs before performing GI analysis and carefully evaluate all results from such custom jobs. Contigs will be ordered against a user-selected reference genome using the Mauve contig orderer (**Rissman et al., 2009**). Please note that annotation of the contigs is still required.

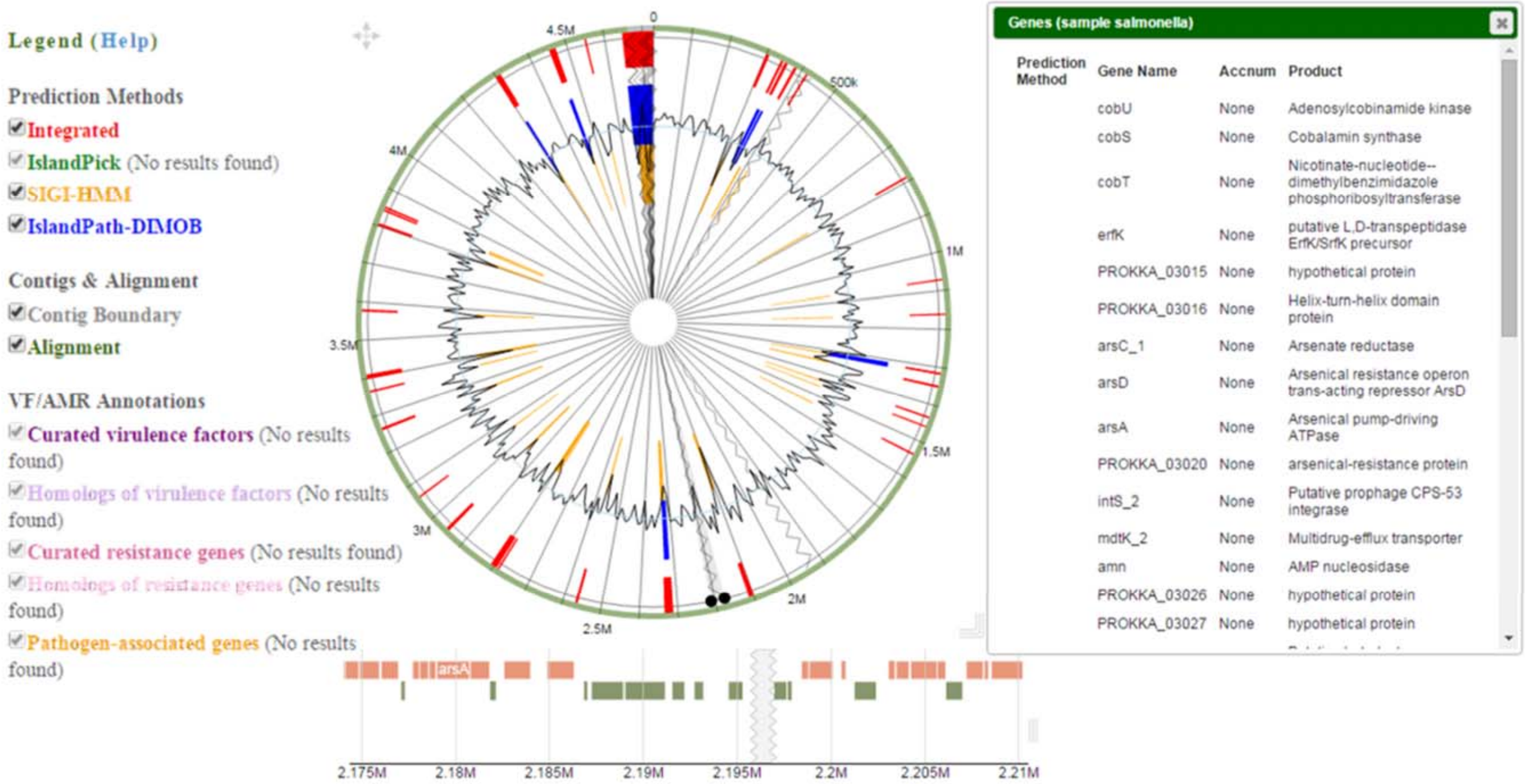
Figure 3.3 Screenshot of the draft genome submission page in IslandViewer 3

When a user attempts to upload a draft genome sequence on the IslandViewer submission page, they are prompted to select a complete reference genome. A screenshot of the submission page upon a user uploading a draft genome is shown in Figure 3.4. Currently, reference genomes must be chosen from pre-computed IslandViewer genomes, which include all NCBI complete microbial genomes as of the most recent IslandViewer update. Multiple users have requested to be able to upload their own reference genome, so this functionality may be added to the draft genome pipeline soon. Users must submit annotated contigs in GenBank or EMBL format.

An earlier version of the draft genome pipeline had included an optional genome annotation step which used Prokka (Seemann, 2014), but this option was not implemented into IslandViewer 3. This option was removed in part to be consistent with the normal IslandViewer custom genome analysis pipelines, which requires an annotated genome as input. Also, all three GI prediction methods in IslandViewer depend on genome annotations, and differences in genome annotations can lead to different results. This is true for all three prediction methods, but IslandPath-DIMOB in particular searches genome annotations for mobility genes. It is the responsibility of the user to choose an annotation pipeline which suits their needs. Automated selection of a reference genome used for contig ordering was also an option in an earlier version of the draft genome pipeline, but this option was also left out of IslandViewer 3, so users are responsible for manually selecting a reference genome. Choice of reference sequence has the potential to impact results, particularly in the case of GIs which overlap with contig edges. The importance of a proper choice of reference sequence has been well characterized for SNV analyses (Pightling et al., 2014).

3.4.3. Draft Genome Visualization

The IslandViewer visualization of draft genome results is similar to the interface for complete genomes (see 3.4.2), but contains a few extra features. An example for results of a draft genome is shown in Figure 3.3. Contigs, which are arranged based on their alignment to the complete reference, are shown in a circular view made using GenomeD3Plot (Laird et al., 2015). Contigs which could not be aligned to the reference are shown at the top of the circular view, just before the origin. Aligned contigs are denoted by a green line around the outside of the circular view which does not cover unaligned contigs. Contig boundaries are shown in the circular view as jagged grey lines. These jagged lines also appear at contig boundaries in the linear view. Like other components of the IslandViewer visualization, contig boundary and alignment labels can be turned on or off by selecting these labels in the legend.



Legend (Help)

Prediction Methods

☑Integrated

☑IslandPick (No results found)

☑SIGI-HMM

☑IslandPath-DIMOB

Contigs & Alignment

☑Contig Boundary

☑Alignment

VF/AMR Annotations

☑Curated virulence factors (No results found)

☑Homologs of virulence factors (No results found)

☑Curated resistance genes (No results found)

☑Homologs of resistance genes (No results found)

☑Pathogen-associated genes (No results found)

Genes (sample salmonella)

Prediction Method	Gene Name	Accnum	Product
	cobU	None	Adenosylcobinamide kinase
	cobS	None	Cobalamin synthase
	cobT	None	Nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase
	erfK	None	putative L,D-transpeptidase ErfK/SrfK precursor
	PROKKA_03015	None	hypothetical protein
	PROKKA_03016	None	Helix-turn-helix domain protein
	arsC_1	None	Arsenate reductase
	arsD	None	Arsenical resistance operon trans-acting repressor ArsD
	arsA	None	Arsenical pump-driving ATPase
	PROKKA_03020	None	arsenical-resistance protein
	intS_2	None	Putative prophage CPS-53 integrase
	mdtK_2	None	Multidrug-efflux transporter
	amn	None	AMP nucleosidase
	PROKKA_03026	None	hypothetical protein
	PROKKA_03027	None	hypothetical protein

Figure 3.4 The IslandViewer interface for a draft genome. This example shows *Salmonella enterica* subsp. enterica serovar Newport str. SL254

3.4.4. Caveats and benefits

The IslandViewer draft genome pipeline was developed due to high demand. With this update, IslandViewer is able to predict GIs in draft genome sequences, which constitute the vast majority of all bacterial and archaeal genome sequences. However, draft genomes are not ideal for GI analysis. The reasons for this are described in detail in Chapter 2. The IslandViewer submission page includes a warning regarding false predictions and missing GIs in draft genomes, and it is recommended that users only submit high quality draft genomes for which a similar reference genome is available.

That being said, in many cases it is currently impractical to complete every genome that is being analyzed, particularly in larger projects which sequence hundreds or thousands of microbial genomes. Chapter 2 also described how IslandViewer was able to identify many true GIs. The majority of genomes submitted to IslandViewer for custom analysis are now drafts. IslandViewer facilitates the analysis of GIs for genomes which previously could not be analyzed and true, valuable GI predictions can be obtained this way.

Chapter 4.

Concluding remarks

The two main goals of my project have been to perform an initial characterization of the importance of missing regions in draft genomes, and to increase the functionality of IslandViewer software to allow for the analysis of draft genomes. Through the work described in this thesis, I have achieved both of these goals. To achieve the first goal, I used two sets of draft and complete genomes from two very different bacterial species: *Listeria monocytogenes* and *Pseudomonas aeruginosa*. Several important results were common between these data sets. In both data sets, neither antimicrobial resistance genes nor virulence factors were disproportionately missing from draft genomes. This result is encouraging for the use of draft genomes for clinical or epidemiological purposes—at least for the species examined. tRNA genes and replication, recombination, and repair genes were both disproportionately missing from draft genomes, which is of particular interest because both of these gene types are associated with genomic islands (GIs). Results of GI analysis differed between the two data sets, but this may be due at least in part to the lack of draft and complete genomes produced from identical isolates in the *P. aeruginosa* data set. In both data sets, GI predictions were missed in the draft genome, and many GIs were predicted at contig boundaries. This is notable, since GIs are noted for potentially encoding genes involved in recent adaptations. These findings form the basis for further study, and are useful for researchers, clinicians, or epidemiologists to consider when interpreting GI predictions in draft genomes.

The second goal was achieved with the release of IslandViewer 3 (Dhillon et al., 2015), which includes a simple pipeline to allow draft genomes as input. While there are limitations to using draft genomes for analysis of GIs, they can still be used to produce valuable results. There are several ways in which the IslandViewer draft genome pipeline may be improved in the near future. These include warnings for genomic regions with potential false negative (such as contig boundaries where a GI is predicted in the similar reference genome) or potential false positive (such as GI predictions influenced by automated gene annotation) predictions. Due to demand, the ability for users to upload a custom reference genome to be used for contig alignment may also be incorporated soon.

A new GI analysis tool, IslandCompare, is being developed by the Brinkman Laboratory in order to directly compare GI predictions in dozens of genomes at once, including draft genomes.

The number of draft bacterial genomes being produced vastly outnumbers complete genomes, and bacterial genomics is being used more extensively in clinical and epidemiological settings. To accommodate this, I have set out to characterize the features and limitations of draft genomes, and to enable GI analysis using draft genomes. Further research on the characteristics of draft genomes, particularly for other species of particular clinical and epidemiological importance, will be key as draft genome analysis becomes more prevalent in these areas.

References

Abreu, A. C., Ramos, R. T. J., Cerdeira, L., Silva, A., Soares, S. C., Trost, E., Tauch, A., Jr, R. H., … Azevedo, V. (2012). PIPS : Pathogenicity Island Prediction Software, *7*(2). doi:10.1371/journal.pone.0030848

Adhya, S. L., & Shapiro, J. A. (1969). The galactose operon of E. coli K-12. I. Structural and pleiotropic mutations of the operon. *Genetics*, *62*(2), 231.

Allard, M. W., Strain, E., Melka, D., Bunning, K., Musser, S. M., Brown, E. W., & Timme, R. (2016). the PRACTICAL value of Food Pathogen Traceability through BUILDING a Whole-Genome Sequencing Network and database. *Journal of Clinical Microbiology*, (March), JCM.00081–16. doi:10.1128/JCM.00081-16

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Alvarez-Martinez, C., & Christie, P. (2009). Biological diversity of prokaryotic type IV secretion systems. *Microbiology and Molecular Biology Reviews : MMBR*, *73*(4), 775–808. doi:10.1128/MMBR.00023-09

Arvey, A. J., Azad, R. K., Raval, A., & Lawrence, J. G. (2009). Detection of genomic islands via segmental genome heterogeneity, *37*(16), 5255–5266. doi:10.1093/nar/gkp576

Athey, T. B. T., Teatero, S., Takamatsu, D., Wasserscheid, J., Dewar, K., Gottschalk, M., & Fittipaldi, N. (2016). Population Structure and Antimicrobial Resistance Profiles of Streptococcus suis Serotype 2 Sequence Type 25 Strains. *PloS One*, *11*(3), e0150908.

Bailly-Bechet, M., Vergassola, M., & Rocha, E. (2007). Causes for the intriguing presence of tRNAs in phages. *Genome Research*, *17*(10), 1486–1495.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., … Prjibelski, A. D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. doi:10.1016/j.cell.2007.05.009

Bashir, A., Klammer, A. A., Robins, W. P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., … Schadt, E. E. (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology*, *30*(7), 701–7. doi:10.1038/nbt.2288

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Grif, S., Khanna, A., Marshall, M., … Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res*, *32*(Database issue), 138D–41. doi:10.1093/nar/gkh121

Bellanger, X., Payot, S., Leblond-Bourget, N., & Guédon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiology Reviews*, *38*(4), 720–760.

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, gks001.

Bi, D., Xu, Z., Harrison, E. M., Tai, C., Wei, Y., He, X., Jia, S., Deng, Z., … Ou, H.-Y. (2011). ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Research*, gkr846.

Blunsom, P. (2004). Hidden markov models. *Lecture Notes, August*, *15*, 18–19.

Bouvier, M., Demarre, G., & Mazel, D. (2005). Integron cassette insertion: a recombination process involving a folded single strand substrate. *The EMBO Journal*, *24*(24), 4356–4367. Retrieved from http://emboj.embopress.org/content/24/24/4356.abstract

Burget, C., Campbello, A. M., Karlint, S., & Campbell, A. M. (1992). Over-and under-representation of short oligonucleotides in DNA sequences. *Genetics*, *89*, 1358–1362.

Cambray Guillaume, Guerout Anne-Marie, & Mazel, D. (2010). Integrons. *Annual Review of Genetics*, *44*(1), 141–166. doi:10.1146/annurev-genet-102209-163504

Campbell, A. M. (1992). Chromosomal insertion sites for phages and plasmids. *Journal of Bacteriology*, *174*(23), 7495–7499.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.-L., & Brüssow, H. (2003). Phage as agents of lateral gene transfer. *Current Opinion in Microbiology*, *6*(4), 417–424.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Molecular Microbiology*, *49*(2), 277–300.

Chain, P. S. G., Grafham, D. V, Fulton, R. S., FitzGerald, M. G., Hostetler, J., Muzny, D., Ali, J., Birren, B., … Detter, J. C. (2009). Genome Project Standards in a New Era of Sequencing. *Science*, *326*(5950), 236–237. doi:10.1126/science.1180614

Chen, L., Xiong, Z., Sun, L., Yang, J., & Jin, Q. (2012). VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Research*, *40*(D1), D641–D645. doi:10.1093/nar/gkr989

Cherf, G. M., Lieberman, K. R., Rashid, H., Lam, C. E., Karplus, K., & Akeson, M. (2012).

Automated Forward and Reverse Ratcheting of DNA in a Nanopore at Five Angstrom Precision. *Nature Biotechnology*, *30*(4), 344–348. doi:10.1038/nbt.2147

Chiapello, H., Gendrault, A., Caron, C., Blum, J., Petit, M.-A., & El Karoui, M. (2008). MOSAIC: an online database dedicated to the comparative genomics of bacterial strains at the intra-species level. *BMC Bioinformatics*, *9*(1), 1–9. doi:10.1186/1471-2105-9-498

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., … Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Meth*, *10*(6), 563–569. Retrieved from 10.1038/nmeth.2474

Clark, S. C., Egan, R., Frazier, P. I., & Wang, Z. (2013). ALE: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, *29*(4), 435–443. doi:10.1093/bioinformatics/bts723

Clarke, J., Wu, H., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*(April), 265–270. doi:10.1038/nnano.2009.12

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research*, *14*(7), 1394–1403. doi:10.1101/gr.2289704

Darmon, E., & Leach, D. R. F. (2014). Bacterial genome instability. *Microbiology and Molecular Biology Reviews : MMBR*, *78*(1), 1–39. doi:10.1128/MMBR.00035-13

de Brito, D. M., Maracaja-Coutinho, V., de Farias, S. T., Batista, L. V, & do Rêgo, T. G. (2016). A Novel Method to Predict Genomic Islands Based on Mean Shift Clustering Algorithm. *PLoS ONE*, *11*(1), e0146352. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pone.0146352

De Soyza, A., Hall, A. J., Mahenthiralingam, E., Drevinek, P., Kaca, W., Drulis□Kawa, Z., Stoitsova, S. R., Toth, V., … Zlosnik, J. E. A. (2013). Developing an international Pseudomonas aeruginosa reference panel. *Microbiologyopen*, *2*(6), 1010–1023.

Delcher, A. L., Harmon, D., Kasif, S., White, O., & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, *27*(23), 4636–4641. doi:10.1093/nar/27.23.4636

Derrington, I. M., Butler, T. Z., Collins, M. D., Manrao, E., Pavlenok, M., Niederweis, M., & Gundlach, J. H. (2010). Nanopore DNA sequencing with MspA. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16060–16065. doi:10.1073/pnas.1001831107

Dhillon, B. K., Chiu, T. A., Laird, M. R., Langille, M. G. I., & Brinkman, F. S. L. (2013). IslandViewer update: improved genomic island discovery and visualization. *Nucleic Acids Research* , *41* (W1), W129–W132. doi:10.1093/nar/gkt394

Dhillon, B. K., Laird, M. R., Shay, J. A., Winsor, G. L., Lo, R., Nizam, F., Pereira, S. K., Waglechner, N., … Brinkman, F. S. L. (2015). IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Research*, *43*(W1), W104–8. doi:10.1093/nar/gkv401

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Micro*, *2*(5), 414–424. Retrieved from http://dx.doi.org/10.1038/nrmicro884

Doolittle, W. F., & Bapteste, E. (2007). Pattern pluralism and the Tree of Life hypothesis. *Proceedings of the National Academy of Sciences*, *104*(7), 2043–2049.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H. O. K., Buffalo, V., … Koren, S. (2011). Assemblathon 1: a competitive assessment ofde novo short read assembly methods. *Genome Res*, *21*, 2224–2241. doi:10.1101/gr.126599.111

Eddy, S. (2003). HMMER User's Guide. Biological sequence analysis using profile hidden Markov models.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput Biol*, *7*(10), e1002195.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., … Turner, S. (2009). Real-time DNA Sequencing from Single Polymerase Molecules. *Exchange Organizational Behavior Teaching Journal*, (January), 133–138.

Eklund, M. W., Poysky, F. T., Reed, S. M., & Smith, C. A. (1971). Bacteriophage and the toxigenicity of Clostridium botulinum type C. *Science*, *172*(3982), 480–482.

Ereshefsky, M. (2010). Microbiology and the species problem. *Biology and Philosophy*, *25*(4), 553–568. doi:10.1007/s10539-010-9211-9

Escudero, J. A., Loot, C., Nivina, A., & Mazel, D. (2015). The integron: adaptation on demand. *Microbiology Spectrum*, *3*(2).

Finkel, S. E., & Kolter, R. (2001). DNA as a nutrient: novel role for bacterial competence gene homologs. *Journal of Bacteriology*, *183*(21), 6288–6293.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, gkr367.

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., … Bateman, A. (2006). Pfam: clans, web tools and services.

Nucleic Acids Research , *34* (suppl 1), D247–D251. doi:10.1093/nar/gkj149

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., … Venter, J. C. (1995). Whole genome random sequencing and assembly of Haemophilus influenzae rd. *Science*, *269*(5223), 496–512. doi:10.1126/science.7542800

Flores, C. O., Meyer, J. R., Valverde, S., Farr, L., & Weitz, J. S. (2011). Statistical structure of host–phage interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(28), E288–E297. doi:10.1073/pnas.1101595108

Flusberg, B. A., Webster, D., Lee, J., Travers, K., Olivares, E., Clark, A., Korlach, J., Turner, S. W., … Park, M. (2010). Direct detection of DNA methylation during single-molecule, real- time sequencing. *Nature Methods*, *7*(6), 461–465. doi:10.1038/nmeth.1459.Direct

Fortier, L.-C., & Sekulovic, O. (2013). Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence*, *4*(5), 354–365.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., … Venter, J. C. (1995). The Minimal Gene Complement of Mycoplasma genitalium. *Science*, *270*(5235), 397–403. doi:10.1126/science.270.5235.397

Freeman, V. J. (1951). Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. *Journal of Bacteriology*, *61*(6), 675.

Freifelder, D., & Meselson, M. (1970). Topological relationship of prophage λ to the bacterial chromosome in lysogenic cells. *Proceedings of the National Academy of Sciences*, *65*(1), 200–205.

Freschi, L., Jeukens, J., Kukavica-Ibrulj, I., Boyle, B., Dupont, M.-J., Laroche, J., Larose, S., Maaroufi, H., … Levesque, R. C. (2015). Clinical utilization of genomics data produced by the international Pseudomonas aeruginosa consortium. *Frontiers in Microbiology*, *6*, 1036. doi:10.3389/fmicb.2015.01036

Fuhrman, J. a., & Noble, R. T. (1995). Viruses and protists cause similar bacterial mortality in coastal seawater. *Limnology and Oceanography*, *40*(7), 1236–1242. doi:10.4319/lo.1995.40.7.1236

Garcia-Vallve, S., Guzman, E., Montero, M. A., & Romeu, A. (2003). HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Research*, *31*(1), 187–189. doi:10.1093/nar/gkg004

Gellatly, S. L., & Hancock, R. E. W. (2013). Pseudomonas aeruginosa: new insights into pathogenesis and host defenses. *Pathogens and Disease*, *67*(3), 159–173.

Georg, J., & Hess, W. R. (2011). cis-antisense RNA, another level of gene regulation in bacteria. *Microbiology and Molecular Biology Reviews : MMBR*, *75*(2), 286–300. doi:10.1128/MMBR.00032-10

Gibson, M. K., Forsberg, K. J., & Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME Journal*, *9*(1), 207–216.

Gilchrist, C. A., Turner, S. D., Riley, M. F., Petri, W. A., & Hewlett, E. L. (2015). Whole-Genome Sequencing in Outbreak Analysis. *Clinical Microbiology Reviews*, *28*(3), 541–563. doi:10.1128/CMR.00075-13

Gilmour, M. W., Graham, M., Reimer, A., & Van Domselaar, G. (2013). Public health genomics and the new molecular epidemiology of bacterial pathogens. *Public Health Genomics*, *16*(1-2), 25–30. doi:10.1159/000342709 [doi]

Gilmour, M. W., Graham, M., Van Domselaar, G., Tyler, S., Kent, H., Trout-Yakel, K. M., Larios, O., Allen, V., … Nadon, C. (2010). High-throughput genome sequencing of two Listeria monocytogenes clinical isolates during a large foodborne outbreak. *BMC Genomics*, *11*(1), 120. doi:10.1186/1471-2164-11-120

Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., Berche, P., Bloecker, H., … Chakraborty, T. (2001). Comparative genomics of Listeria species. *Science*, *294*(5543), 849–852.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, *17*(6), 333–351. Retrieved from http://dx.doi.org/10.1038/nrg.2016.49

Gosalbes, M. J., Durbán, A., Pignatelli, M., Abellan, J. J., Jiménez-Hernández, N., Pérez-Cobas, A. E., Latorre, A., & Moya, A. (2011). Metatranscriptomic approach to analyze the functional human gut microbiota. *PloS One*, *6*(3), e17447.

Griffith, F. (1928). The Significance of Pneumococcal Types. *Epidemiology & Infection*, *27*(02), 113–159.

Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Lopez-Rojas, R., Kempf, M., Landraud, L., & Rolain, J.-M. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial Agents and Chemotherapy*, *58*(1), 212–220.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. doi:10.1093/bioinformatics/btt086

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R., & 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. *Microbial Pathogenesis*, *8*(3), 213–225.

Hamon, M., Bierne, H., & Cossart, P. (2006). Listeria monocytogenes: a multifaceted model. *Nature Reviews Microbiology*, *4*(6), 423–434.

Hao, W., & Golding, G. B. (2006). The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Research*, *16*(5), 636–643.

Hernandez, D., François, P., Farinelli, L., Østerås, M., & Schrenzel, J. (2008). De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Research*, *18*(5), 802–809. doi:10.1101/gr.072033.107

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., & Zamir, A. (1965). Structure of a ribonucleic acid. *Science*, *147*(3664), 1462–1465.

Hsiao, W. W. L., Ung, K., Aeschliman, D., Bryan, J., Finlay, B. B., & Brinkman, F. S. L. (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet*, *1*(5), e62.

Hsiao, W., Wan, I., Jones, S. J., & Brinkman, F. S. L. (2003). IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, *19*(3), 418–420. doi:10.1093/bioinformatics/btg004

Hudson, C. M., Lau, B. Y., & Williams, K. P. (2015). Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes. *Nucleic Acids Research*, *43*(D1), D48–D53. doi:10.1093/nar/gku1072

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., & Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biology*. doi:10.1186/gb-2013-14-5-r47

Hutchison, C. A. (2007). DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Research*, *35*(18), 6227–6237. doi:10.1093/nar/gkm688

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119. doi:10.1186/1471-2105-11-119

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M. B., Pope, B. J., Tomita, T., Zobel, J., Holt, K. E., … Howden, B. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine*, *6*(11), 90. doi:10.1186/s13073-014-0090-6

Ip, C. L. C., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., Leggett, R. M., Eccles, D. A., … Consortium, M. A. and R. (2015). MinION Analysis and Reference

Consortium: Phase 1 data release and analysis. *F1000Research*, *4*, 1075. doi:10.12688/f1000research.7201.1

Jeukens, J., Boyle, B., Kukavica-Ibrulj, I., Ouellet, M. M., Aaron, S. D., & Fleiszig, S. (2014). Comparative Genomics of Isolates of a Pseudomonas aeruginosa Epidemic Strain.

Johnston, C., Martin, B., Fichant, G., Polard, P., & Claverys, J.-P. (2014). Bacterial transformation: distribution, shared mechanisms and divergent control. *Nature Reviews. Microbiology*, *12*(3), 181–96. doi:10.1038/nrmicro3199

Jorgensen, J. H., Ferraro, M. J., Jorgensen, J. H., & Ferraro, M. J. (2009). Antimicrobial susceptibility testing: a review of general principles and contemporary practices. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *49*(11), 1749–55. doi:10.1086/647952

Karaolis, D. K. R., Johnson, J. A., Bailey, C. C., Boedeker, E. C., Kaper, J. B., & Reeves, P. R. (1998). A Vibrio cholerae pathogenicity island associated with epidemic and pandemic strains. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(6), 3134–3139. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC19707/

Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13770–13773. doi:10.1073/pnas.93.24.13770

Klassen, J., & Currie, C. (2012). Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics*, *13*(1), 14. Retrieved from http://www.biomedcentral.com/1471-2164/13/14

Klockgether, J., Cramer, N., Wiehlmann, L., Davenport, C. F., & Tümmler, B. (2011). Pseudomonas aeruginosa genomic structure and diversity. *Pseudomonas Aeruginosa, Biology, Genetics, and Host-Pathogen Interactions*, 6.

Koren, S., Treangen, T., Hill, C., Pop, M., & Phillippy, A. (2014). Automated ensemble assembly and validation of microbial genomes. *BMC Bioinformatics*, *15*(1), 126. Retrieved from http://www.biomedcentral.com/1471-2105/15/126

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., … Hattori, M. (2007). Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes. *DNA Research*, *14*(4), 169–181. doi:10.1093/dnares/dsm018

Laird, M. R., Langille, M. G. I., & Brinkman, F. S. L. (2015). GenomeD3Plot: A library for rich, interactive visualizations of genomic data in web applications. *Bioinformatics*, *31*(20), 3348–3349. doi:10.1093/bioinformatics/btv376

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., … International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. doi:10.1038/35057062

Langille, M. G. I. (2009). *Computational prediction and characterization of genomic islands: insights into bacterial pathogenicity*. *Department of Molecular Biology and Biochemistry*. Simon Fraser University. Retrieved from http://morganlangille.com/papers/Langille_2009_Thesis.pdf

Langille, M. G. I., & Brinkman, F. S. L. (2009). IslandViewer: An integrated interface for computational identification and visualization of genomic islands. *Bioinformatics*, *25*(5), 664–665. doi:10.1093/bioinformatics/btp030

Langille, M. G. I., Hsiao, W. W., & Brinkman, F. S. (2008a). Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics*, *9*(1), 329. doi:10.1186/1471-2105-9-329

Langille, M. G. I., Hsiao, W. W. L., & Brinkman, F. S. L. (2010). Detecting genomic islands using bioinformatics approaches. *Nat Rev Micro*, *8*(5), 373–382. Retrieved from http://dx.doi.org/10.1038/nrmicro2350

Langille, M. G. I., Laird, M. R., Hsiao, W. W. L., Chiu, T. A., Eisen, J. A., & Brinkman, F. S. L. (2012). MicrobeDB: a locally maintainable database of microbial genomic sequences. *Bioinformatics*, *28*(14), 1947–1948. doi:10.1093/bioinformatics/bts273

Langille, M. G. I., Zhou, F., Fedynak, A., Hsiao, W. L., Xu, Y., & Brinkman, F. S. L. (2008b). Mobile Genetic Elements and Their Prediction. *Computational Methods for Understanding Bacterial and Archaeal Genomes*, *7*, 113.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), 1–10. doi:10.1186/gb-2009-10-3-r25

Laslett, D., & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, *32*(1), 11–16. doi:10.1093/nar/gkh152

Laszlo, A. H., Derrington, I. M., Ross, B. C., Brinkerhoff, H., Adey, A., Nova, I. C., Craig, J. M., Langford, K. W., … Gundlach, J. H. (2014). Decoding long nanopore sequencing reads of natural DNA. *Nature Biotechnology*, *32*(8), 829–833. doi:10.1038/nbt.2950

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

Lee, C.-C., Chen, Y.-P. P., Yao, T.-J., Ma, C.-Y., Lo, W.-C., Lyu, P.-C., & Tang, C. Y.

(2013). GI-POP: A combinational annotation and genomic island prediction pipeline for ongoing microbial genome projects. *Gene*, *518*(1), 114–123. doi:10.1016/j.gene.2012.11.063

Liu, B., & Pop, M. (2009). ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Research*, *37*(Database), D443–D447. doi:10.1093/nar/gkn656

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Meth*, *12*(8), 733–735. Retrieved from 10.1038/nmeth.3444

Loman, N. J., & Watson, M. (2015). Successful test launch for nanopore sequencing. *Nat Meth*, *12*(4), 303–304. Retrieved from http://dx.doi.org/10.1038/nmeth.3327

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*(5), 955–964. doi:10.1093/nar/25.5.955

Lwoff, A. (1953). Lysogeny. *Bacteriological Reviews*, *17*(4), 269.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L. J., & Salzberg, S. L. (2013). GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, *29*(14), 1718–1725. doi:10.1093/bioinformatics/btt273

Mahillon, J., & Chandler, M. (1998). Insertion sequences. *Microbiology and Molecular Biology Reviews*, *62*(3), 725–774.

Mamanova, L., Andrews, R. M., James, K. D., Sheridan, E. M., Ellis, P. D., Langford, C. F., Ost, T. W. B., Collins, J. E., & Turner, D. J. (2010). FRT-seq: Amplification-free, strand-specific, transcriptome sequencing. *Nature Methods*, *7*(2), 130–132. doi:10.1038/nmeth.1417

Marchler-Bauer, A., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., … Bryant, S. H. (2009). CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Research* , *37* (suppl 1), D205–D210. doi:10.1093/nar/gkn845

Mardis, E. R. (2011). A decade’s perspective on DNA sequencing technology. *Nature*, *470*(7333), 198–203. doi:10.1038/nature09796

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. a, Berka, J., Braverman, M. S., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–80. doi:10.1038/nature03959

Marvel, C. C. (1986). A program for the identification of tRNA-like structures in DNA sequence data. *Nucleic Acids Research*, *14*(1), 431–435.

Maxam, a M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560–564. doi:10.1073/pnas.74.2.560

Mazel, D., Dychinco, B., Webb, V. A., & Davies, J. (1998). A Distinctive Class of Integron in the Vibrio cholerae Genome. *Science*, *280*(5363), 605–608. Retrieved from http://science.sciencemag.org/content/280/5363/605.abstract

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., … Wright, G. D. (2013). The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, *57*(7), 3348–3357. doi:10.1128/AAC.00419-13

McClintock, B. (1941). The stability of broken ends of chromosomes in Zea mays. *Genetics*, *26*(2), 234–282.

Metzker, M. L. M. L. L. (2005). Emerging technologies in DNA sequencing. *Genome Res.*, *15*(12), 1767–76. doi:10.1101/gr.3770505.with

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, *35*(Web Server issue), W182–W185. doi:10.1093/nar/gkm321

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., … Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*. doi:10.1093/nar/gkr344

Nyrén, P., Pettersson, B., & Uhlén, M. (1993). Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical Biochemistry*. doi:10.1006/abio.1993.1024

O'Brien, A. D., Newland, J. W., Miller, S. F., Holmes, R. K., Smith, H. W., & Formal, S. B. (1984). Shiga-like toxin-converting phages from Escherichia coli strains that cause hemorrhagic colitis or infantile diarrhea. *Science*, *226*(4675), 694–696.

Ou, H.-Y., He, X., Harrison, E. M., Kulasekara, B. R., Thani, A. Bin, Kadioglu, A., Lory, S., Hinton, J. C. D., … Deng, Z. (2007). MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands. *Nucleic Acids Research*, *35*(suppl 2), W97–W104.

Oxford Nanopore. (n.d.). Start using MinION. Retrieved May 17, 2016, from https://www.nanoporetech.com/community/start-using-minion

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., … Milos, P. M. (2009). Direct RNA sequencing. *Nature*, *461*(7265), 814–818. Retrieved from http://dx.doi.org/10.1038/nature08390

Pacific Biosciences. (2015). Revolutionize Genomics with SMRT Sequencing. Retrieved May 20, 2016, from http://www.pacb.com/wp-content/uploads/2015/09/Revolutionize-Genomics-with-SMRT-Sequencing.pdf

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*. doi:10.1101/gr.186072.114

Paul, J. H. (2008). Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J*, *2*(6), 579–589. Retrieved from http://dx.doi.org/10.1038/ismej.2008.35

Phillippy, A. M., Schatz, M. C., & Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol*, *9*(3), R55. doi:10.1186/gb-2008-9-3-r55

Pightling, A. W., Petronella, N., & 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*(8), e104579. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pone.0104579

Podell, S., & Gaasterland, T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biology*, *8*(2), 1.

Pundhir, S., Vijayvargiya, H., & Kumar, A. (2008). PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes. *In Silico Biology*, *8*(3, 4), 223–234.

Rahman, A., & Pachter, L. (2013). CGAL: computing genome assembly likelihoods. *Genome Biol*, *14*(1), R8. doi:10.1186/gb-2013-14-1-r8

Rajan, I., Aravamuthan, S., & Mande, S. S. (2007). Identification of compositionally distinct regions in genomes using the centroid method. *Bioinformatics*, *23*(20), 2672–2677.

Redfield, R. J. (1988). Evolution of bacterial transformation: is sex with dead cells ever better than no sex at all? *Genetics*, *119*(1), 213–221.

Ribeiro, F. J., Przybylski, D., Yin, S., Sharpe, T., Gnerre, S., Abouelleil, A., Berlin, A. M., Montmayeur, A., … Jaffe, D. B. (2012). Finished bacterial genomes from shotgun sequence data. *Genome Research*, *22*(11), 2270–2277. doi:10.1101/gr.141515.112

Rissman, A. I., Mau, B., Biehl, B. S., Darling, A. E., Glasner, J. D., & Perna, N. T. (2009). Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics*, *25*(16), 2071–2073. doi:10.1093/bioinformatics/btp356

Roach, J. C., Boysen, C., Wang, K., & Hood, L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, *26*(2), 345–353.

Roberts, A. P., Chandler, M., Courvalin, P., Guédon, G., Mullany, P., Pembroke, T., Rood, J. I., Jeffery Smith, C., … Berg, D. E. (2008). Revised nomenclature for transposable genetic elements. *Plasmid*, *60*(3), 167–173. doi:http://dx.doi.org/10.1016/j.plasmid.2008.08.001

Rocha, E. P. C., & Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends in Genetics*, *18*(6), 291–294. doi:10.1016/S0168-9525(02)02690-2

Roche Applied Science. (2011). Sanger-like read lengths - the power of next-gen throughput Uncover more of the genome through extended coverage. *Www.My454.Com*, *June*, 1–4.

Rohwer, F., & Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of Bacteriology*, *184*(16), 4529–4535.

Rowe, W., Baker, K. S., Verner-Jeffreys, D., Baker-Austin, C., Ryan, J. J., Maskell, D., & Pearce, G. (2015). Search engine for antimicrobial resistance: a cloud compatible pipeline and web interface for rapidly detecting antimicrobial resistance genes directly from sequence data. *PloS One*, *10*(7), e0133492.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., … Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, *22*(3), 557–567. doi:10.1101/gr.131383.111

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, a R., Fiddes, C. a, Hutchison, C. a, Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, *265*(5596), 687–695. doi:5,386 base pairs...

Sanger, F., & Nicklen, S. (1977). DNA sequencing with chain-terminating. *Pnas*, *74*(12), 5463–5467. doi:http://dx.doi.org/10.1073%2Fpnas.74.12.5463

Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Research*, *20*(9), 1165–1173. doi:10.1101/gr.101360.109

Schoner, B., Geistlich, M., Rosteck, P., Rao, R. N., Seno, E., Reynolds, P., Cox, K., Burgett, S., & Hershberger, C. (1992). Sequence similarity between macrolide-resistance determinants and ATP-binding transport proteins. *Gene*, *115*(1-2), 93–96. doi:10.1016/0378-1119(92)90545-Z

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. doi:10.1093/bioinformatics/btu153

Shapiro, J. A. (1969). Mutations caused by the insertion of genetic material into the galactose operon of Escherichia coli. *Journal of Molecular Biology*, *40*(1), 93–105.

Shapiro, J. A., & Adhya, S. L. (1969). The galactose operon of E. coli K-12. II. A deletion analysis of operon structure and polarity. *Genetics*, *62*(2), 249.

Shrivastava Sakshi, Ch V Siva Kumar Reddy, S. S. M. (2010). INDeGenIUS , a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. *Journal of Biosciences*, *35*(3), 351–364. doi:10.1007/s12038-010-0040-4

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V, & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. doi:10.1093/bioinformatics/btv351

Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A., & Alm, E. J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, *480*(7376), 241–244.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B., & Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, *321*(6071), 674–679. doi:10.1038/321674a0

Soares, S. C., Geyik, H., Ramos, R. T. J., de Sá, P. H. C. G., Barbosa, E. G. V, Baumbach, J., Figueiredo, H. C. P., Miyoshi, A., … Silva, A. (2015). GIPSy: genomic island prediction software. *Journal of Biotechnology*.

Stokes, H. W. t, & Hall, R. M. (1989). A novel family of potentially mobile DNA elements encoding site□specific gene□integration functions: integrons. *Molecular Microbiology*, *3*(12), 1669–1683.

Stover, C. K., Pham, X. Q., Erwin, a L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., … Olson, M. V. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, *406*(6799), 959–964. doi:10.1038/35023079

Sueoka, N. (1992). Directional Mutation Pressure, Selective Constraints, and Genetic Equilibria. *J Mol Evol*, *34*, 95–114.

Sui, S. J. H., Fedynak, A., Hsiao, W. W. L., Langille, M. G. I., & Brinkman, F. S. L. (2009). The association of virulence factors with genomic islands. *PloS One*, *4*(12), e8094.

Suttle, C. A. (2005). Viruses in the sea. *Nature*, *437*(7057), 356–361. Retrieved from http://dx.doi.org/10.1038/nature04160

Suttle, C. A. (2007). Marine viruses — major players in the global ecosystem. *Nat Rev Micro*, *5*(10), 801–812. Retrieved from http://dx.doi.org/10.1038/nrmicro1750

Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids*

Research, *28*(1), 33–36.

Tatusov, R. L., Koonin, E. V, & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, *278*(5338), 631–637. doi:10.1126/science.278.5338.631

Thayer, A. M. (2014). Next -Gen Sequencing Is A Numbers Game. *Chemical & Engineering News*, *92*(33), 11–15.

Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, *15*(11), 524. doi:10.1186/PREACCEPT-2573980311437212

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46. doi:10.1038/nrg3117

Tsai, I. J., Hunt, M., Holroyd, N., Huckvale, T., Berriman, M., & Kikuchi, T. (2014). Summarizing specific profiles in illumina sequencing from whole-genome amplified DNA. *DNA Research*, *21*(3), 243–254. doi:10.1093/dnares/dst054

Tu, Q., & Ding, D. (2003). Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS Microbiology Letters*, *221*(2), 269–275. Retrieved from http://femsle.oxfordjournals.org/content/221/2/269.abstract

Turcatti, G., Romieu, A., Fedurco, M., & Tairi, A. P. (2008). A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research*, *36*(4). doi:10.1093/nar/gkn021

van Schaik, W., Top, J., Riley, D. R., Boekhorst, J., Vrijenhoek, J. E. P., Schapendonk, C. M. E., Hendrickx, A. P. A., Nijman, I. J., … Tettelin, H. (2010). Pyrosequencing-based comparative genome analysis of the nosocomial pathogen Enterococcus faecium and identification of a large transferable pathogenicity island. *BMC Genomics*, *11*(1), 1.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. doi:10.1126/science.1058040

Vernikos, G. S., & Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics* , *22* (18), 2196–2203. doi:10.1093/bioinformatics/btl369

Vincent, A. T., Derome, N., Boyle, B., Culley, A. I., & Charette, S. J. (2016). Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. *Journal of Microbiological Methods*, (April).

doi:10.1016/j.mimet.2016.02.016

Vold, B. S. (1985). Structure and organization of genes for transfer ribonucleic acid in Bacillus subtilis. *Microbiological Reviews*, *49*(1), 71.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W., Surovcik, K., Meinicke, P., … Lawrence, J. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*, *7*(1), 142. doi:10.1186/1471-2105-7-142

Wagner, E. G. H., & Simons, R. W. (1994). Antisense RNA Control in Bacteria, Phages, and Plasmids. *Annu. Rev. Microbiol.*, *48*, 713–42. doi:10.1146

Wang, Q., Holmes, N., Martinez, E., Howard, P., Hill-Cawthorne, G., & Sintchenko, V. (2015). It is not all about SNPs: comparison of mobile genetic elements and deletions in Listeria monocytogenes genomes links cases of hospital-acquired listeriosis to the environmental source. *Journal of Clinical Microbiology*, JCM–00202.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., … Sobral, B. W. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research*, *42*(D1), D581–D591. doi:10.1093/nar/gkt1099

Wei, W., Gao, F., Du, M., Hua, H., Wang, J., & Guo, F. (2016). Zisland Explorer : detect genomic islands by combining homogeneity and heterogeneity properties, (November 2015), 1–10. doi:10.1093/bib/bbw019

Whiteley, M., Bangera, M. G., Bumgarner, R. E., Parsek, M. R., Teitzel, G. M., Lory, S., & Greenberg, E. P. (2001). Gene expression in Pseudomonas aeruginosa biofilms. *Nature*, *413*(6858), 860–864.

Williams, D., Paterson, S., Brockhurst, M. A., & Winstanley, C. (2016). Refined analyses suggest that recombination is a minor source of genomic diversity in Pseudomonas aeruginosa chronic cystic fibrosis infections. *Microbial Genomics*, *2*(3).

Woese, C. (1998). The universal ancestor. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(12), 6854–6859. doi:10.1073/pnas.95.12.6854

Wozniak, R. A. F., & Waldor, M. K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews. Microbiology*, *8*(8), 552–63. doi:10.1038/nrmicro2382

Wu, J. Q., Du, J., Rozowsky, J., Zhang, Z., Urban, A. E., Euskirchen, G., Weissman, S., Gerstein, M., & Snyder, M. (2008). Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. *Genome Biology*, *9*(1), 1–14. doi:10.1186/gb-2008-9-1-r3

Xu, Z., & Hao, B. (2009). CVTree update: A newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Research*, *37*(SUPPL. 2), 174–178. doi:10.1093/nar/gkp278

Yoon, S. H., Park, Y.-K., & Kim, J. F. (2014). PAIDB v2. 0: exploration and analysis of pathogenicity and resistance islands. *Nucleic Acids Research*, gku985.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *The Journal of Antimicrobial Chemotherapy*, *67*(11), 2640–4. doi:10.1093/jac/dks261

Zhang, X., McDaniel, A. D., Wolf, L. E., Keusch, G. T., Waldor, M. K., & Acheson, D. W. K. (2000). Quinolone antibiotics induce Shiga toxin-encoding bacteriophages, toxin production, and death in mice. *Journal of Infectious Diseases*, *181*(2), 664–670.

Zhou, C. E., Smith, J., Lam, M., Zemla, A., Dyer, M. D., & Slezak, T. (2007). MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Research*, *35*(suppl 1), D391–D394.