PRINCIPLES FOR DEVELOPING ROBUST CORE GENOME MULTILOCUS SEQUENCE TYPING SYSTEMS

DILLON O. R. BARKER
Bachelor of Science, University of Manitoba, 2014

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfillment of the
Requirements for the Degree

MASTER OF SCIENCE

Department of Biological Sciences
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

PRINCIPLES FOR DEVELOPING ROBUST CORE GENOME MULTILOCUS SEQUENCE TYPING SYSTEMS

DILLON O. R. BARKER

Date of Defense: April 13, 2018

Dr. J. E. Thomas Co-Supervisor	Professor Emeritus	Ph.D.
Dr. E. N. Taboada Co-Supervisor	Adjunct Professor	Ph.D.
Dr. V. P. J. Gannon Committee Member	Adjunct Professor	Ph.D.
Dr. C. R. Laing Committee Member	Research Scientist	Ph.D.
Dr. T. M. Burg Committee Member	Associate Professor	Ph.D.
Dr. I. Kovalchuk Examination Chair	Associate Professor	Ph.D.

Abstract

Core genome multilocus sequence typing is a next generation typing method for the long-term tracking of pathogenic bacteria. Although such methods provide the very high discriminatory power required by public health agencies, they are prone to difficulties relating to data loss intrinsic to current DNA sequencing technologies.

This thesis describes a framework for developing conservative, but powerful core genome multilocus sequencing systems. To this end, I developed a prototype scheme for *Campylobacter jejuni* consisting of 697 core genome loci identified through the analysis of 5,693 *C. jejuni* whole genome sequences. I surveyed the extent of missing data in the dataset, and studied optimizing number of genes to include in such a scheme. Using the information learned in the survey of missing data, I developed a system for predicting unknown alleles from core genome typing data. The principles learned through my research can be applied to develop robust methods of pathogen surveillance.

Acknowledgments

No thesis is written in a vacuum. I'd like to thank all the people who helped me with the research and writing that went into this. Of course I'd like to thank my supervisors, Ed and Jim, and the rest of my committee for their supervision of this process, and especially for the revisions they suggested. Often they were challenging, but all of them improved my work. I would also like to thank Viviana Lartiga and Dr. Robert Wood for their invaluable assistance in helping me at the 11th hour.

Thank you Matt for talking over any number of projects, ideas, and technologies. Ben and Cassandra, we've had some great adventures over the years. Cody and Ruth, you've been amazing friends this whole time. Peter and Steven, thanks for setting me on this whole adventure. Emily, thanks for *everything.*

Contents

List of Tables

List of Figures

List of Abbreviations

AWC	Adjusted Wallace Coefficient
BLAST	Basic Local Alignment Search Tool
bp	base pair
CC	Clonal Complex
CGF	Comparative Genomic Fingerprinting
cgMLST	Core Genome Multilocus Sequence Typing
DNA	Deoxyribonucleic acid
ENA	European Nucleotide Archive
JSON	JavaScript Object Notation
MALDI-TOF	Matrix-assisted Laser Desorption/Ionization Time-of-Flight
MEE	*see MLEE*
MLEE	Multilocus Enzyme Electrophoresis
MLST	Multilocus Sequence Typing
MLVA	Multiple Locus VNTR Analysis; *see VNTR*
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PFGE	Pulsed-field Gel Electrophoresis
RFLP	Restriction Fragment Length Polymorphism

RNA	Ribonucleic acid
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
spp	Species
ST	Sequence Type
UN	United Nations
VNTR	Variable Number of Tandem Repeats
WGS	Whole Genome Sequencing

Chapter 1

A Review of Current Literature

1.1 Molecular Typing and Public Health

In the aftermath of a devastating earthquake, Haiti experienced an outbreak of cholera beginning in October 2010. *Vibrio cholerae* — the bacterium responsible for cholera — had not been recorded in the small Caribbean nation for nearly a century. At the time of writing, the epidemic is ongoing and has claimed the lives of nearly 10,000 Haitians and has spread into the neighbouring Dominican Republic.

In October of 2010, journalists noted the unsanitary conditions in a military encampment inhabited by United Nations peacekeepers in the Artibonite River Valley in rural Haiti. Notably, a pipe drained untreated sewage from the camp into the river. To relieve Bangladeshi troops, Nepalese peacekeepers rotated into the area in early October.

Soon after the arrival of the Nepalese troops, Haitian public health officials noted a sharp increase in dysenteric illness. Laboratory assays identified *V. cholerae* serogroup O1, serotype Ogawa, biotype El Tor as the causative agent of the outbreak. Retrospective analysis of hospital records revealed cholera cases beginning inland near the UN camp on 17 October and spreading downstream along the Artibonite River until reaching the coast on 22 October [1].

Though epidemiological evidence implicated the UN troops, other hypotheses, such as an increase in temperature and salinity had allowed endemic *Vibrio* to proliferate, had not been excluded. In 2011, Hendriksen *et al.* characterized *V. cholerae* strains isolated from confirmed Haitian cholera cases along with strains isolated from a concurrent epidemic in Nepal. This study showed that the Haitian cholera strains were closely related to one another and shared a recent single common ancestor. Moreover, the group sharing the recent common ancestor in the Haitian isolates also included *V. cholerae* strains isolated in Bangladesh and Nepal. A 2013 study by Katz *et al.* using whole genome sequencing confirmed these results. The evidence was deemed to be strong support of the hypothesis that cholera was introduced inadvertently by Nepalese UN peacekeepers deployed to the area [2–4].

Though this investigation came too late to prevent the Haiti outbreak, it was an effective illustration of the power of molecular typing and whole genome sequencing in determining the source of a catastrophic epidemic.

Public health is, at its core, the science of improving health at the scale of populations. Public health agencies take a wide variety of approaches to reduce the burden of illness upon their citizens. Infectious diarrhœal illness caused by pathogenic bacteria are amongst the largest sources of loss of disability-adjusted life years for all ages and sexes in both developing and advanced economies [5].

In the context of epidemiologic investigations, molecular typing is the differentiation of microbial strains on the basis of differences at the molecular level [6]. These differences may be between expressed amino acid sequences, the nucleotide sequences that encode them, or even non-coding genetic regions. These differences may be detected indirectly, such as through immunologic reactivity or oligonucleotide hybridization, or directly through DNA sequencing. The two main measures of a typing system are typability and discriminatory power. Typability is the reliability with which a type can be assigned to a subject organism, and discriminatory power is the capacity for a typing system to distinguish between two similar strains.

Effective public health interventions rely on accurate and timely identification of microbial isolates. Molecular typing data can provide the discriminatory power necessary to answer key questions about the strains of interest. In the abstract, all such questions ask, *"Is this strain the same as that strain?"* More concretely, we can derive practical information like whether a given strain is included in an outbreak or is part of the background of sporadic cases, or the likelihood of a particular isolate having originated from a particular source.

1.2 Biochemical & Antigenic Typing Methods

1.2.1 Biotyping

Biotyping includes a broad spectrum of typing systems that compare biochemical differences amongst bacterial isolates. Biotyping methods focus on colony morphology; chemical resistances, including antibiotic sensitivity/resistance patterns; environmental resistances; and isolate metabolic processes, such as substrate catabolism and metabolites produced. Biotyping methods are able to provide discriminatory power ranging from the genus level to the subspecies level, depending upon the organism and panel of tests employed. To be useful, these traits must vary significantly amongst the organism to be typed [7–9].

Biotyping methods are generally fast, technically forgiving, and inexpensive to perform. Typability is usually very high. These features make biotyping attractive for processing large numbers of strains and, despite the problems discussed below, biotyping is often sufficient for identification of bacterial isolates to the species level.

Unfortunately, classical biotyping methods are moderately reproducible at best, and, unless a large number of traits are investigated, suffer from poor discriminatory power [7–9]. Additionally, biotyping has a demonstrated capacity to lead investigators to incorrect conclusions as to the identity of the organism in question. Maslow *et al.* described a case in which two separate *Klebsiella* isolates were drawn from the same patient and biotyped as *K. pneumoniae* and *K. oxytoca.* Later analysis showed that the two isolates belonged to the same clone and differed only in their production of indole [10]. Similarly, the biotyping scheme for *Campylobacter* distinguishes between *Campylobacter jejuni* and *Campylobacter coli* on the basis of their respective positive and negative hippurate production [11]. However, it has since been shown that some strains of *C. jejuni* are hippurate-negative, and that differentiation between these species based on this trait is not always supported by genetic evidence [12].

Though it was largely rendered obsolete by later advances in microbial typing, biotyping is being modernized through large, high-throughput phenotypic assays. Often automated, these modern cousins of classical biotyping have many aspects in common, and can assess dozens-to-hundreds of phenotypic traits, particularly growth substrates and antibiotic resistances. These systems have been successfully used for a variety of purposes, ranging from serotype and virulence prediction to national public health surveillance programmes [13–15].

1.2.2 Lysis Typing: Bacteriophages & Bacteriocins

Amongst the earliest typing systems were two methods different in origin, but similar in interpretation: bacteriophage, or simply 'phage' typing, and bacteriocin typing. Both methods operate on the variable and binary nature of susceptibility to the inhibitory agent in question.

Phages that have host ranges below the species level and are obligately lytic are candidates for use in a phage typing system. The bacterial strain being studied is co-incubated with different variants of the typing phage, and sensitivity is observed as plaques on a bacterial lawn, or clearing if a liquid medium is used [16].

Bacteriocins are toxic proteins and small peptides produced by some bacteria that have extremely narrow spectra of target strains, and a high specific activity. These toxins are employed by a bacterium to kill closely related strains for a competitive advantage [17]. As with phage typing, a panel of representative bacteriocins is added to a lawn of the strain being tested, and growth inhibition, if any, is noted after incubation

[18–20].

Phage typing was first developed by Craigie and Yen for *Salmonella enterica* subspecies *enterica* serovar Typhi[*] [16]. In their experiments with the Type II Vi phage, they made two important observations: the phage often reacted weakly, or not at all, with a given *S. enterica* Typhi strain; and if phage particles were isolated after weakly reacting and added to a fresh culture of the same *S. enterica* strain, an aggressively lytic reaction would be observed. These features were exploited to create a standardized panel of phage types that could be used in a binary typing scheme. The Type II Vi phage was co-incubated with each of a set of reference strains. Subsequent strains being phage typed were assigned a phage type based on the pattern of sensitivity to the adapted reference phages.

Bacteriocin typing has a similar history, beginning with the work of Abbott and Shannon in 1957. The investigators developed a typing system based upon the inhibition patterns of *Shigella sonnei* by seven variants of the bacteriocin colicine. In their pilot study, the authors were able to group 367 of 537 *S. sonnei* strains into seven colicine types, with the balance untypable. This study laid the groundwork for later bacteriocin typing systems [18]. Some later schemes combined phage typing with bacteriocin typing into a single assay. In combination, the increased number of possible types improves discrimination with little additional technical challenge [19].

The principal advantage of lysis typing is its quick turnaround time, which allows large numbers of isolates to be processed quickly, making it a valuable technique for reference laboratories [7, 8]. However, phage typing is considered to be very technically demanding. The need to cultivate extensive libraries of standardized phage cultures also keeps this typing method practical only for large reference laboratories [7, 8]. Phage typing and bacteriocin typing have poor discriminatory power when compared to modern typing systems, though this can be ameliorated somewhat by using

[*] The authors use the now-obsolete name *Bacillus typhosus*

them in conjunction with one another [7, 8, 19].

1.2.3 Serotyping

Serotyping is based on the differential reaction between known antibodies and unknown proteinaceous or carbohydrous antigens on the surface of a bacterial cell [21–23]. The specific pattern of agglutination reactions between a panel of known antibodies and an isolate form the serotype, which is synonymously referred to as the serovar. In the event an isolate does not react to any of the antibodies of a given serotyping scheme, it is first considered untypable, though it may prove to be a candidate for a novel serotype.

Serotyping was first described as a technique by Lancefield in 1933, which she developed during her study of human- and food-associated *Staphylococcus haemolyticus*. The method was later adapted to many other bacteria, notably *Salmonella enterica*, *Escherichia coli*, and *Campylobacter* species. The Kauffmann-White scheme for *S. enterica* and classification of *E. coli* by their O- and H- antigens continue to be of particular importance to the modern terminology for these organisms [22, 24, 25].

Though it revolutionized bacterial typing, traditional agglutination-based serotyping is not without its disadvantages. It is exceptionally demanding of a technician's time, labour, and skills. Moreover, the monoclonal antibodies comprising the antisera are difficult and expensive to produce [7, 26, 27].

1.3 Polymerase Chain Reaction-based Methods

Typing methods based upon the polymerase chain reaction (PCR) are many and varied. Three broad categories of PCR-based methods are discussed here: analysis of variable numbers of tandem repeats, random amplification, and binary presence/absence surveys.

1.3.1 Variable Number of Tandem Repeats

Within genomes, there regions where short, repetitive patterns of nucleotides called tandem repeats are known to exist. Analysis of the variable number of tandem repeats (VNTR) uses this number as a characteristic fingerprint of the strain. During bacterial chromosome replication, these regions are prone to slipped strand mispairing, which can lead to the gain or loss of these repeat units [28]. The number of repeats can be inferred from amplicon mobility following electrophoresis. VNTR analysis can be enhanced by using multiple target loci, and is known as multiple locus VNTR analysis, or MLVA. Because this method yields relatively high resolution, is inexpensive, and is easy to perform and analyze; MLVA was once considered to be a potential 'gold standard' assay for molecular typing of certain pathogenic bacteria, including *Staphylococcus aureus* and *Mycobacterium tuberculosis* [29–31]. However, because tandem repeats can evolve quickly, the rate of change in these regions may outpace the overall evolution of the strain, sometimes giving incongruous relationships between strains [8].

1.3.2 Random Amplification

Random amplification PCR uses a solitary short primer pair of arbitrary sequence. When amplified under low-stringency conditions, a banding pattern that is characteristic of the genome appears [32, 33]. Visualized by gel electrophoresis, random amplification can provide relatively high discriminatory power, surpassing that of Multilocus Enzyme Electrophoresis, which will be discussed in Section 1.5.1 [34]. Random amplification also benefits from being a quick and inexpensive procedure to perform. However, results cannot be easily compared between laboratories, as they fluctuate and are sensitive to small variations between technicians, reagents, and hardware [6].

1.3.3 Binary Typing

PCR can be used to query for the existence of a particular locus, or at least the existence of its primer binding sites. When this is applied to a panel of genes, a characteristic pattern of locus presence/absence can be described. When selecting target loci for a binary PCR typing system, two general approaches may be taken.

The first is to prioritize selection of loci that are predictive of the organism's epidemicity, pathogenicity, or other features of interest. P-BIT, and its successor method MBiT, both embody this philosphy of binary PCR typing [35, 36].

The second approach to developing a binary PCR typing system is to select markers on the basis of discriminatory power. Typifying this approach is Comparative Genomic Fingerprinting (CGF), which has been developed for use in *C. jejuni*, *E. coli*, and *Arcobacter butzleri* [37–40]. In the *C. jejuni* scheme, a panel of forty target genes were selected on the premise of their approximately 50% carriage in the population. The profiles generated lend themselves to hierarchical clustering, and the method itself is rapid and low-cost. More importantly, CGF produces epidemiologically useful clusters and profiles are readily portable between laboratories [37, 38].

1.4 Restriction Fragment Length Polymorphism

Restriction fragment length polymorphism (RFLP) methods assess diversity within a species by using restriction endonucleases to cut DNA into smaller, variably-sized fragments. The frequency with which a restriction enzyme digests the subject DNA is governed by the length of its recognition site; short recognition sites will cut more frequently and produce shorter DNA fragments than long recognition sites. These chromosomal segments are electrophoresed and the diversity of the resultant banding patterns are used as a fingerprint with which to compare different isolates.

Ribotyping is a variant of RFLP which uses relatively frequent-cutting (4 bp target) enzymes to cut ribosomal DNA. Following digestion and electrophoresis, Southern blot hybridization is used to clarify polymorphisms within ribosomal operons.

Though ribotyping is easily outmatched in terms of discriminatory power by its contemporaries and modern methods alike, its limited diversity of signal was also a strength in the context of outbreak investigations. Two strains of the same outbreak almost surely had identical ribotypes, and so having differing signals would likely indicate that two strains were not closely related. Outside of outbreak scenarios, ribotyping often has limited utility for distinguishing between members of a single species [8].

Pulsed-field gel electrophoresis (PFGE) is a RFLP technique which combines infrequently-cutting restriction enzymes ($\geq$ 6 bp target) with an alternating electric field, as opposed to the constant electric field used in most electrophoretic methods. By alternating the polarity of the electric field, PFGE is better able to resolve subtle differences in mobility of large chromosomal DNA molecules than other electrophoretic methods. By resolving these differences, PFGE is able to make use of a greater range of fragment sizes than other methods, and is thus more effective at distinguishing between similar strains [41].

Pulsed-field gel electrophoresis was developed by Schwartz and Cantor in 1984 to overcome the inability of previous gel electrophoresis methods to adequately resolve large DNA fragments (*i.e.* >50 kilobases) [41]. Though the authors developed PFGE with the intention of karyotyping yeast, the method was later adapted to incorporate restriction endonucleases, as is the case in other RFLP methods. The combination of very high resolution and readily comparable electrophoretic band patterns lead to the adoption of PFGE as the 'gold standard' typing method for many different bacteria. In 1995, the United States Centers for Disease Control and Prevention in conjunction with a number of state-level public health laboratories implemented PulseNet, a PFGE-based national surveillance programme for *E. coli* O157:H7, nontyphoidal *Salmonella*, *Listeria monocytogenes*, and *Shigella* [42]. The PulseNet protocol was later exported internationally and expanded to other organisms [43].

In contrast to ribotyping, PFGE exhibits very high resolution between strains, has been successfully used to characterize bacterial strains within an outbreak, and yields reproducible fingerprints for routine surveillance that can be easily shared between labs [42–44].

As with all gel electrophoresis methods, restriction endonuclease based typing is both costly and challenging. PFGE in particular is well known for long turnaround times and the need for careful analysis [7, 8].

1.5 Allele Typing

1.5.1 Multilocus Enzyme Electrophoresis

Multilocus Enzyme Electrophoresis (MEE or MLEE) is a molecular typing technique which exploits variability in the degree of electrophoretic mobility for a collection of hydrophilic intracellular housekeeping enzymes [45]. Non-synonymous mutations in the underlying gene change the amino acid sequences of the enzymes, and thus alter their molecular weight and net electrostatic charge. After being electrophoresed on a cold potato starch gel, enzyme mobilities are visualized by adding the relevant substrate to each. Coloured products generated by enzymatic catabolism of the substrates indicates the position of each enzyme. The specific rate of travel for each enzyme is its electromorph. Each unique combination of individual electromorphs is known as an electromorph type [45].

MLEE was first developed in 1966 for studying the population structure of *Drosophila pseudoobscura*, and separately, the polymorphism of blood enzymes in *Homo sapiens* [46, 47]. The method later found exploratory use as a typing system for pathogenic bacteria, pioneered in *E. coli* by Caugant, Ochman, Achtman, and their respective colleagues [48–50]. When compared to preceding typing methods, MLEE offered high discrimination between strains. In particular, MLEE was successfully used to discover diversity within serotypes and to characterize population structure in *E. coli* and *Helicobacter pylori*, amongst other bacterial species [51–53].

While MLEE was a powerful tool in the past for investing microbial diversity and population structure, it inherits the difficulties of any gel electrophoresis-based method: it is slow to perform and requires the labour and care of a skilled laboratory technician to generate reproducible results. Post-transcriptional modification of target enzymes can further complicate interpretation of MLEE data, and is considered a source of error [54, 55]. Finally, an electromorph may be degenerate for several underlying alleles whose translation products have indistinguishable mobilities [45]. Together, these factors prevented MLEE from being used in clinical settings or for outbreak investigations [56].

The most important legacy of MLEE was to lay the conceptual groundwork for the later nucleotide-based system of multilocus sequence typing, which quickly superseded it [8, 57].

1.5.2 Single Locus Sequence Typing

Single Locus Sequence Typing involves the analysis of a single highly variable gene or gene region within the organism of interest. The locus of interest is amplified by PCR before Sanger sequencing [27, 58, 59]. Once the nucleotide sequence has been determined, a multiple sequence alignment of all investigated variants of the locus is performed, and pairwise distances are calculated [60].

Two historically important intraspecies single locus sequence typing schemes were *emm* typing of *Streptococcus pyogenes*, and *fla* typing of *C. jejuni*. Each investigates a hypervariable region of their namesake gene. Occasionally, single gene schemes, such as *porA* typing for *C. jejuni*, were used to enhance the resolving power of more recently developed multiple locus typing systems (see below) [61].

On a grander scale, the gene encoding the 16S small ribosomal subunit shared by all prokaryotic life has been used to establish phylogenetic relationships. Originally characterized by the banding given by digestion with T1 RNase (see ribotyping above), Woese and Fox studied 16S ribosomal RNA to discover Archaea and establish our current understanding of the three domain system [62]. Later, researchers used the nucleotide sequences of the 16S ribosomal DNA to identify and infer relationships amongst bacteria. This type of analysis was facilitated by storage of 16S sequences in curated publicly accessible databases [63, 64].

Single locus sequence typing methods were often able to place organisms into epidemiologically or phylogenetically useful groups [58, 60, 61]. Amongst sequence-based typing methods, these are arguably the simplest to perform.

Later multiple locus methods categorically eclipsed their single locus antecedents, excepting their occasional use as an additional enhancing locus. These multiple locus methods were only incrementally more difficult, but offered a much higher resolution alternative. In some cases, it was possible for single hypervariable genes to mutate faster than the actual spread of a pathogen. In an outbreak investigation, this could distort the apparent number of sources [8].

1.5.3 Multilocus Sequence Typing

Multilocus sequence typing (MLST) considers the allelic diversity of a small number — typically five to ten — ‘housekeeping’ genes. These housekeeping genes carry out functions essential to cell survival, and thus evolve slowly and exhibit universal carriage within a species. In MLST, each novel allele is assigned a number corresponding to the order of its discovery and characterization, *i.e.* allele 1 of a target gene was its first described variant, allele 2 its second, and so on. Alleles were generally determined by Sanger sequencing of the target loci [59]. Typically, loci are approximately 500 bp long regions within the target genes, flanked by highly conserved primer binding sites. The definition of each allele is subject to manual curation and submitted to and stored in a centralized database, thereby guaranteeing that a given allele name always refers to the same underlying nucleotide sequence, and *vice versa* [57]. Perhaps the largest such database is PubMLST, maintained by the University of Oxford (http://pubmlst.org) [65]. Each unique combination of alleles is considered a Sequence Type (ST), and related STs may be further grouped into Clonal Complexes (CC). Analysis and clustering of MLST results is straightforward; the pairwise Hamming distance of allele calls at each target gene, *i.e.* the number of differences between two allele profiles, is taken as the phylogenetic distance between two strains [66].

MLST was published by Maiden *et al.* in 1998, with a pilot study conducted using *Neisseria meningitidis*. This prototype scheme consisted of six loci ranging in length from 433 to 501 bp [57]. This general methodology was later applied to other organisms, and there are currently 125 different MLST schemes hosted on PubMLST [65]. The core idea of using a small number of housekeeping genes was adapted from MLEE. While MLEE attempts to infer the allele from changes in electrophoretic mobility stemming from changes in peptide charge or length, MLST interrogates the underlying nucleotide sequence. The use of housekeeping genes was essential to the design of MLST; besides ensuring their presence, and thereby the typability of the strain, the slow evolution of these genes made MLST an appropriate tool for studying the long term evolution of the population structure of a species on a global scale [57, 67].

The principal advantage of MLST is portability. Many earlier methods suffered from poor reproducibility within a laboratory, or lacked a means of sharing data in such a way that the assigned type meant the same thing irrespective of time, location, or interpretation. MLST holds particular advantage for analysis of highly recombinogenic organisms. Because any genetic change will define a new allele, instances of both vertically-inherited point mutations and horizontal homologous recombination are abstracted as equivalent genetic events. Without this consideration, a recombina-

Table 1.1: Reagent costs and turnaround time from pure culture for selected molecular typing methods.

Method	Reagent Cost	Turnaround Time	Citation
CGF	$6.75	5 h	[70]
MALDI-TOF	$0.50	5.1 m	[71]
MLEE	€6	Several days	[72, 73]
MLST	€18–50	9 d	[6, 74]
MLVA	€8	3 h	[6, 75]
PFGE	€20	24–30 h	[6, 76, 77]
Phage Typing	$10	15–18 h	[78, 79]
Serotyping	$15.30–42.79	2–3 d	[80, 81]

tion event can distort apparent distance by instantaneously introducing a large number of pairwise nucleotide differences relative to a strain's closest neighbour [68].

The use of housekeeping genes makes MLST a largely inappropriate choice for outbreak or short term epidemiological investigation [6, 8, 69]. STs change too slowly to reflect evolution within the short time frame of an outbreak. However, MLST can be used to provide evidence that a strain at least belongs to an outbreak, as opposed to a coincidental sporadic case, as outbreak members are likely to share a ST [38]. Because generation of the allelic profiles is generally performed via Sanger sequencing, MLST can be a costly and laborious affair [6, 8, 59].

MLST remains popular for genetic analysis of bacterial populations. Since the invention of the original MLST schemes, there has been interest in extending the MLST concept to greater numbers of genes in the pursuit of enhanced discriminatory power. The advent of inexpensive whole genome sequencing has driven development of MLST-like systems which attempt to target all genes which exhibit universal carriage within a species. Such efforts are discussed in greater depth in Section 1.6.3.

1.6 Genomics and Proteomics

1.6.1 Matrix-assisted Laser Desorption/Ionization

Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) is a mass spectrometry technique which has recently found use for the typing of pathogenic microorganisms. In its simplest form, MALDI-TOF works on a sample of raw, unprocessed bacterial cells, either placed directly on the target plate, or in liquid suspension. Some protocols instead use a solution of extracted proteins rather than whole cells [82]. The biological sample is vapourized by a laser beam and passed through a powerful electric field. Analogously to mobilities observed during an agarose gel electrophoresis, small ions arrive at the detector more quickly than large ones. Complex spectra are generated, which may be used as characteristic fingerprints of a particular bacterial type [8, 82, 83].

Perhaps MALDI-TOF's greatest advantage is its extraordinarily fast turnaround time from sample to answer. A fingerprint can be generated in minutes using this method. In clinical settings or in the midst of an outbreak, where time is of the essence, this advantage cannot be overstated. Because very small quantities are required of the biological input, MALDI-TOF can avoid selection bias in cases where the act of culturing a microbe distorts the apparent diversity of a sample [83]. The method may also be used to determine the presence or absence of bacterial toxins and antibiotic resistance factors in the sample [84].

Although MALDI-TOF can very rapidly identify microbial samples in a clinical setting, its utility is limited for identification below the species level, and lags significantly behind contemporary methods with respect to resolution. In *E. coli* and *S. enterica*, MALDI-TOF has been used to successfully determine serotype. In *Pseudomonas putida*, *Streptococcus pyogenes*, *Streptococcus agalactiae*, and other bacterial species, MALDI-TOF achieves discriminatory power similar to single locus typing methods, such as 16S or *gyrB* discussed above [84]. Additionally, while the reagent cost per isolate is on the order of one dollar, the capital cost of the apparatus is hundreds of thousands of dollars [71, 84].

Today, MALDI-TOF is used primarily for identifying the species and serotype of a sample in a clinical setting. It is not widely employed for subtyping below the species level.

1.6.2 Single Nucleotide Polymorphism Typing

Single nucleotide polymorphism (SNP) typing is the categorization of an organism based on the observed nucleotide at specific positions along the chromosome [85]. SNP typing only investigates relatively rare polymorphisms resulting from vertically-inherited mutations, and so nucleotide diversity arising from homologous recombination is not generally considered. Due to this, SNP typing is generally only used in organisms with low recombination rates. Many SNP typing methods compare subject genomes against one or more reference genomes [85, 86].

Due to its very high discriminatory power, SNP typing is an effective means of distinguishing between strains with limited genetic or genomic diversity [87]. Because SNPs may be assigned definite locations within the genome, SNP typing methods permit easy interchange of data between laboratories. In some cases, SNPs are of phenotypic or epidemiologic relevance. In *S. aureus* and other species, a point mutation in the DNA gyrase subunit *gyrA* gene can impart resistance to ciprofloxacin and other quinolone-class antibiotics [88].

SNP typing relies upon the availability of high-quality reference genomes. In cases where these are not available, incomplete draft genomes may be used as a substitute, though extra care must be taken to exclude genome sequence regions known to be of low quality from the analysis [87]. Homologous recombination events have the potential to import a large number of SNPs simultaneously, and may increase the apparent distance between closely related strains when that relationship is measured using SNP typing [68]. One approach to limit the effect of homologous recombination on SNP phylogenies is to ignore cases where multiple SNPs are found within a certain distance of one another. For example, SNVPhyl ignores instances of two or more SNPs within a sliding window [89].

SNP typing is widely used for typing of highly clonal organisms, particularly for outbreak investigations. Outbreak strains often have too little genetic diversity to be resolved by typical molecular surveillance methods such as PFGE. In these scenarios, SNP typing may distinguish between strains. Modern SNP typing methods employ whole genome sequencing for both SNP discovery and SNP calling [90].

1.6.3 Genome-scale Multilocus Sequence Typing

Ribosomal MLST, or rMLST, was an early effort to extend the MLST concept beyond the original scheme size of approximately seven loci. This derivative method of MLST targets 53 ribosomal *rps* genes. Because these genes are shared by all bacteria, rMLST aims to be a universal system for classifying bacterial species [67, 91].

As whole genome DNA sequencing becomes increasingly accessible, interest has grown in two genome-scale extensions of the MLST approach: whole genome MLST (wgMLST), which considers every gene available to a target organism; and core genome MLST (cgMLST), which restricts itself to only those genes shared by all members of the species. As their names imply, these new approaches increase the number of target loci from fewer than ten to hundreds or thousands. The key difference between these two systems is how faithfully they adhere to the original MLST concepts. A cgMLST scheme may be seen as a direct extension of classical MLST, while wgMLST deliberately deviates from that pattern. The inclusion of target genes which are not conserved in all members of the species increases discriminatory power, but can complicate interpretation and analysis. One example of these challenges is the case of a typing locus that appears to be absent. It can be difficult to determine whether that locus does not appear in sequencing data because it is absent from the organism's genome, or it was merely lost in one of the many gaps found in draft genome sequences.

Prototype cgMLST schemes have been developed for several species. An international consortium led by Institut Pasteur developed a wgMLST system targeting *L. monocytogenes* for population biology and public health surveillance purposes [92]. Cody, Maiden, and colleagues at Oxford University have developed low-stringency cgMLST and wgMLST schemes that jointly target *C. jejuni* and *C. coli* [93, 94]. Development of robust cgMLST schemes that preserve the principles of classical MLST is an area of active research. Current challenges in cgMLST design concern stable definitions of the core genome and the loss of allele data due to the limitations of genome sequencing technology.

1.7 Bacterial Genomics

One of the key fields of study in modern biology is that of the genome. The genome is the collection of all genetically encoded information within a single organism. Every organism is the expression of its genome.

All of the above-described typing methods in some way exploit or reveal some information about the target bacterial genome. An early example of this is the use of 16S ribosomal DNA to infer phylogenetic relationships between clades of prokaryotes, which helped develop a tree of life for bacteria and archaea [62].

Concrete observations of phenotype led us to genetics, and our aggregate knowledge of genetics in turn led to genomics. As an increasing number of genomes were studied in depth, a new field — pangenomics — has emerged.

1.7.1 The Bacterial Pangenome

The pangenome is a concept that describes the sum of all genes available to a particular group of organisms, *i.e.* one might speak generically of the *Campylobacter* pangenome or specifically *C. jejuni* pangenome. It is the collective genome of a population. The pangenome may itself be divided into two categories: a core genome composed of all the genes found in every group member, and an accessory or 'dispensable' genome consisting of all the genes that are not. Core genes are definitional to a species, and many core genes are essential to survival [95, 96]. The fraction of an individual cell's genome that belongs to the core genome is variable between species. A large majority of genes in a *C. jejuni* or *S. agalactiae* cell are core genes, whereas the genome of an *E. coli* cell has only a minority of core genes [95, 97, 98].

Selective pressure is exerted on genome content. The limiting factor on bacterial growth rates, and as a consequence their fitness, is the time it takes to replicate the chromosome [99]. As additional genes enlargen the chromosome, they slow replication. As such, these accessory genes must provide an adaptive advantage to justify their carriage.

One of the oldest methods for generating a bacterial pangenome involves an all-*versus*-all comparison of all genetic elements, typically using BLAST to determine homology [100–102]. While this family of methods will effectively cluster homologous genes, it suffers from algorithmic complexity, and can become extraordinarily demanding on CPU and memory resources as the number of genomes increases [103].

PANSEQ was written to assess the question of pangenome definition while striving to avoid the additional complexity that arises when sequence data are treated as a series of genes. PANSEQ first aligns all query sequences using MUMMER [104]. Having aligned the input genomes, they are each divided into *k*-length fragments. These fragments are treated as the basic elements of the pangenome. Fragment homology is compared using BLAST, and the presence or absence of a particular fragment in a given genome is determined [105].

While PANSEQ is agnostic to the biological role of the nucleotide sequence, ROARY uses open reading frames as the fundamental units of the pangenome. ROARY takes gene annotations created by annotation software such as PROKKA as input [106]. Homology searches are coordinated by ROARY using a combination of BLASTP and CD-HIT [100, 103, 107]. To resolve a common difficulty encountered in pangenomics, ROARY uses CD-HIT to consolidate paralogous genes to a single representative. It then uses this information to, upon the users preference, treat gene paralogues as alleles of one another, treat them as discrete genes, or to exclude them from the pangenome entirely [103, 107].

As can been seen, there are a variety of approaches to calculating a bacterial organism's pangenome. The most important consideration, besides accuracy, is the pragmatic requirement that the algorithm finish in a timely manner. Methods such as PanOCT and PGAP were valuable tools for working on a small number of genomes, but given that modern draft genome datasets often are comprised of hundreds or thousands of individuals, tools like PANSEQ or ROARY are now essential.

Methods like MLST and its derivatives — particularly cgMLST— have their foundation in a well described pangenome. Because MLST is predicated on the idea of allelic variation in genes shared by all members, inclusion of any accessory genes in such a scheme will lead to spurious pairwise distances, and give the appearance of poor data quality when actual biological absence is truly to blame.

1.8 *Campylobacter jejuni*: A Testbed for cgMLST Design

The population structure of a bacterial species can determine which WGS-based typing approach is most appropriate. SNP typing in theory provides the highest possible resolution between strains, but is susceptible to distortion through mass import of SNPs during a single recombination event. This makes SNP typing useful primarily in highly clonal organisms with low recombination rates, such as *L. monocytogenes* or *Bacillus anthracis*. Conversely and as discussed above, cgMLST and other sequence typing systems treat vertical mutation and horizontal recombination as equivalent events within a locus, and are useful in highly recombinogenic organisms.

The population structure of *C. jejuni* makes it a particularly suitable subject for cgMLST development and analysis. The species is weakly clonal, has high homologous recombination rates, and many strains are naturally competent [108]. This aspect of its biology is reflected in the molecular typing systems currently in use for public health surveillance of *C. jejuni*: MLST and *fla* typing discussed on pages 11 – 14 consider variation on the level of alleles rather than SNPs, and CGF compares *C. jejuni* strains only on the presence or absence of accessory genes (p. 8). Continuing this evolution from a single locus sequencing typing system to a multilocus sequence typing system, the next step is to develop a core genome MLST scheme.

C. jejuni is a small, motile Gram-negative bacterium in the class Epsilonproteobacteria [109]. These bacteria are thermophilic, microaerophilic, and pathogenic in humans. Symptomatic of *C. jejuni* infection is watery diarrhœa with low mortality. In rare cases, Guillain-Barré Syndrome, a rapid paralysis of the peripheral nervous system, may follow infection.

C. jejuni resides in a wide variety of hosts and environments, including domestic and wild birds, cattle, aquatic ecosystems, pigs, and sheep. It is a zoonotic pathogen, and the disease caused by infection of a human host is called campylobacteriosis. Human infection occurs via the fæcal-oral route, typically following contact with contaminated animals or animal products, particularly chicken and cattle [110, 111].

Besides being a good subject for cgMLST development in the technical sense, improved typing and public health surveillance of *C. jejuni* may yield real dividends in the form of prevention through a better understanding of transmission dynamics and source attribution.

Campylobacteriosis is the leading cause of bacterial diarrhœal gastroenteritis worldwide [110, 112]. It is exceedingly prevalent — the reported annual incidence in England and Wales is 105/100,000 people, though rural areas can have much high incidence rates [110]. In New Zealand, annual incidences as high as 578/100,000 in small children and 470/100,000 in adults have been reported [113]. Morbidity is known to be greater in males than females [110, 113]. Campylobacteriosis is believed to be a widely underreported disease, and true incidence rates may be significantly higher than is recorded [110, 112]. Setting aside human misery, and looking instead from an economic perspective, *C. jejuni* is the cause of an enormous drag on human productivity. *Campylobacter* costs the United Kingdom's National Health Service £50 million per year in direct costs of treatment [114]. Each case of *Campylobacter*-associated Guillain-Barré Syndrome costs hundreds of thousands of dollars to treat [115]. The United States spends tens of billions of dollars every year on medical costs and productivity lost to absenteeism as a result of *Campylobacter* infection [116].

Effective surveillance is essential to any coherent public health effort. A detailed view into the population dynamics of a pathogen such as *C. jejuni* is key to predicting its behaviour and preventing its spread. Surveillance efforts rely on accurate information, and modern programmes use molecular typing data to this end. As the cost of whole genome sequencing continues to fall, typing methods which interrogate genome sequence data become increasingly viable. As interest grows in typing methodologies such as cgMLST, *C. jejuni* is emerging as the ideal candidate for the development of such a scheme due to its recombinogenic population and open genome [95, 117].

1.9 An Overview of this Thesis

Core genome multilocus sequence typing schemes have seen active development for several different bacterial species, including *C. jejuni*; however, such schemes are in their early days, and are encountering challenges unforeseen from experience with classical MLST. In particular, where classical MLST required complete sequence data at a fixed set of prescribed loci, many extant cgMLST schemes have comparatively looser requirements. A scheme for *Campylobacter* spp. published by Cody *et al.* avoided incomplete nucleotide sequence data by using *ad hoc* subsets of 1,667 loci [93]. A related 1,343 locus cgMLST system published by the same research group called for a fixed set of loci to be analyzed, though allowing for up to a 5% absence rate for their definition of core genes [94]. Allowing incomplete typing data in any of these ways prevents the unambiguous assignment of a nomenclature, which was one of the principal advantages of classical MLST for public health surveillance programmes.

This thesis addresses some of the problematic aspects of current cgMLST systems. Taken together, the research presented here is a set of rules and remedies for developing robust cgMLST schemes.

The first objective of this thesis is to design a prototype cgMLST scheme. Due to its recombination-prone genome, frequency of analysis by classical MLST, and multiple competing cgMLST against which to compare, *C. jejuni* is an ideal model organism for cgMLST development. To create a robust cgMLST scheme for *C. jejuni* that minimizes systemic biases, the scheme will be defined with as much genomic and provenantial diversity as possible. The scheme will also only be defined for *C. jejuni*. Although it is included in other schemes, *C. coli* will be excluded from this analysis, as will other *Campylobacter* species [93, 94]. By excluding other non-*C. jejuni Campylobacter* species, a more stable cgMLST scheme can be created. This allows the inclusion of the full core genome of *C. jejuni* and does not limit the scheme to the *Campylobacter* core genome, which is intersect of the core genomes of each of its component species.

Having defined a cgMLST scheme, it is important to assess its performance along two key criteria: the number of missing or untypable loci, and discriminatory power. These are used together to test the efficiency of locus inclusion in the scheme. Efficiency is important in cgMLST design as every locus that is included increases discriminatory power for the scheme, it also raises the probability of an error and thus difficulty of unambiguous assignment to a nomenclature system.

Though the research in this thesis proposes methods to mitigate the risks of missing data presented by the inclusion of each locus, having missing loci is inevitable given the large numbers of draft genome sequences involved. If missing data cannot be prevented in these cases, the research here suggests that it may be reversed. A combination of three sources of data tangent to the missing allele call may be used to predict the identity of the missing locus: the allele possessed by the subject genome's closest relative, the relative abundances of alleles in the population, and matching any partially recovered sequence data to known alleles.

By mitigating the problems stemming from data loss, a stable and unambiguous nomenclature becomes a possibility. A robust nomenclature system allows cgMLST to maximize the key benefits of genome-scale MLST while retaining the key benefits of classical MLST: portability between laboratories and the ability to monitor evolution over time.

Chapter 2

Systematic Design of Core Genome Multilocus Sequence Typing Schemes

2.1 Introduction

Effective public health control of *C. jejuni* relies upon the ability to infer phylogenetic relationships between strains through the use of various molecular typing systems. Multilocus sequence typing (MLST) has been one of the most common such typing methods employed by public health surveillance programmes targeting *C. jejuni*. MLST considers the allelic profile of internal gene fragments of seven conserved housekeeping genes. These genes belong to the *C. jejuni* core genome, and are thus known to be present in all members of the species [118]. The nucleotide sequence of each MLST locus is determined and an allele designation is assigned [57]. The Hamming distance of MLST calls may be used to compare strains on a pairwise basis. Modern advances in DNA sequencing technology have made it feasible to use much larger portions of the genome when designing a molecular typing scheme such as MLST. Core genome MLST (CGMLST) is a modern extension of the MLST concept from seven genes to hundreds or thousands in an attempt to exploit as much of the core genome as possible. Increasing the number of genes in this way dramatically increases the capacity of MLST-like systems to distinguish between similar microbial strains [67].

The high resolution of CGMLST when compared to previous systems, such as the classical seven gene MLST scheme, is important in resolving subtle differences between highly similar strains, key to tasks including microbial source tracking, routine surveillance, and outbreak detection. Developing a cgMLST scheme is a non-trivial task, and care must be taken when selecting both the genomes and genes used to define the system. Because MLST profiles are not defined when loci are absent, it is important to be accurate but conservative when determining inclusion/exclusion criteria. Using a set of genes composing the core genome which are known to be present in all individuals of a species allows a cgMLST scheme to remain usable between projects and across time.

2.2 Methods

2.2.1 Dataset Definition & Assembly

All available *C. jejuni* strains as of 2016-11-23 (n = 7,126) were downloaded from the Sequence Read Archive (SRA) and the European Nucleotide Archive (ENA) using fastq-dump, and the resultant files were split into their forward and reverse FASTQ components [119]. Once downloaded, all genomes were assembled using the INNUca short read assembly pipeline, structured on SPAdes 3.9.0 [120, 121]. As a component of the INNUca pipeline, pilon 'polished' the assemblies, improving assembly quality by fixing mis-assembled sequence and filling gaps in sequence data [122]. Kraken and its MiniKraken database was used to identify non-*C. jejuni* sequence data, and remove it from assemblies [123]. Assemblies that were greater than 2.0 Mbp or less than 1.4 Mbp pairs were removed from further analysis (Figure 2.1).

2.2.2 Annotation & Pangenome Description

Open reading frame prediction and gene annotations were performed using prokka 1.12 [106]. These annotated genes were provided to roary 3.7.1 to calculate a pangenome for *C. jejuni* [103]. The core genome here is defined as the set comprising those genes which were found to be present in at least 99.9% of the 5,693 strains that survived

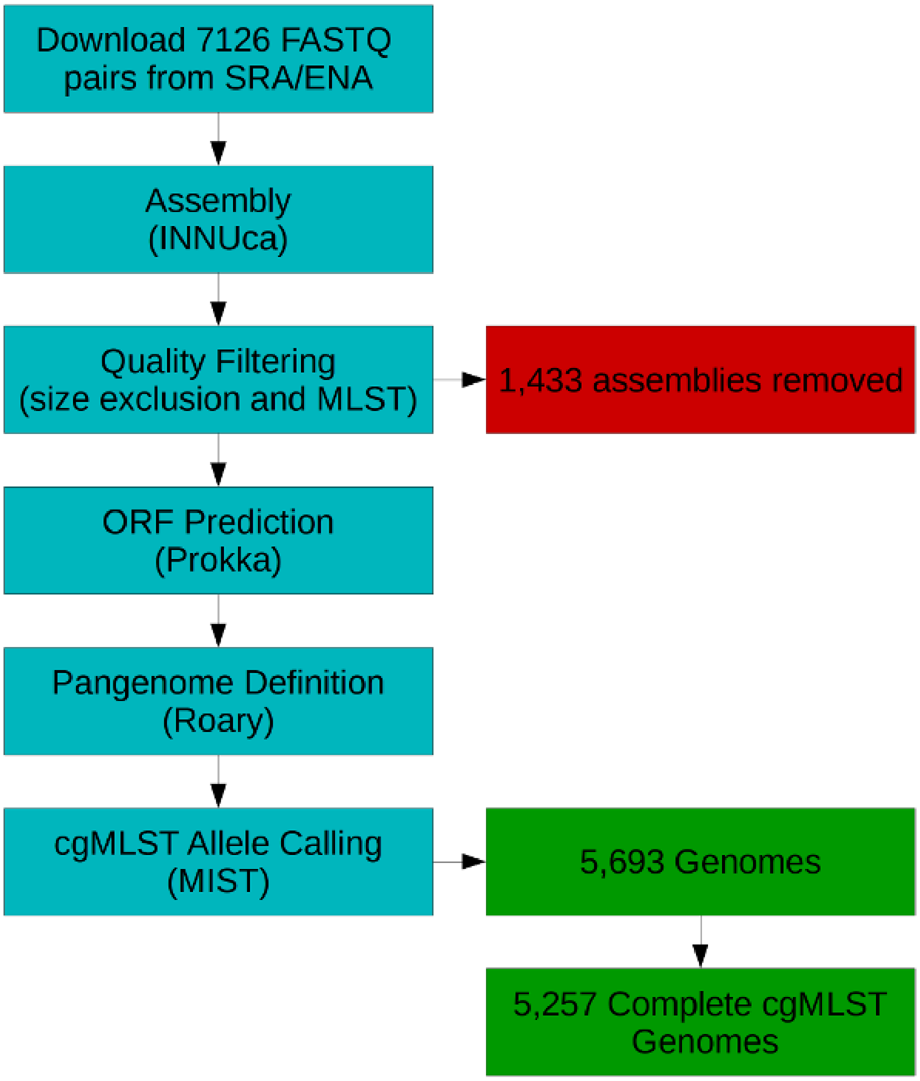
Download 7126 FASTQ pairs from SRA/ENA
Assembly (INNUca)
Quality Filtering (size exclusion and MLST)
1,433 assemblies removed
ORF Prediction (Prokka)
Pangenome Definition (Roary)
cgMLST Allele Calling (MIST)
5,693 Genomes
5,257 Complete cgMLST Genomes

Figure 2.1: Flowchart describing downloading 7,126 genomes from the Sequence Read Archive, developing a 697 locus cgMLST scheme, and extracting high quality core genome profile sets of 5,693 genomes, and a set 5,257 genomes with full typability.

quality and size filtering.

2.2.3 Core Genome Multilocus Sequence Typing Scheme

A representative sequence of each gene identified by ROARY was taken from ERR083867, and named 'allele 1' for each locus. Using the Microbial In Silico Typer (MIST), these representatives were queried against all other assemblies in the data set [124]. Each time a genome in the set of 5,693 genomes that passed quality filtering possessed a previously unobserved allele, this new allele was added to the multifasta of allele definitions for that gene, and assigned an allele number designation. When tabulating the allele calls, '0' represents cases where the query gene could not be found within the subject, and '-1' was used to indicate the presence of a sequencing truncation, *i.e.*, cases where the query was partially found at the end of a contiguous sequence of DNA, or "contig". The processes of assigning allele numbers to novel alleles and tabulating MIST's output data were assisted by two custom Python scripts: update_definitions.py and json2csv.py. Unique cgMLST profiles were extracted using AWK [125]. These scripts are available online at:

`https://github.com/dorbarker/thesis_supporting_scripts`.

The complete workflow from downloading the initial genomes through to defining the cgMLST scheme is depicted in Figure 2.1.

2.3 Results

From the starting 7,126 *C. jejuni* draft genomes, a total of 5,693 genomes were of sufficient quality to not be excluded in Section 2.2.1 and were included in the analysis (Figure 2.2). This became the final set on which subsequent analyses were performed. Although creating this set involved removing poor-quality genomes and genes, it still had low levels of absent and truncated loci. Within this 5,693 strain cgMLST profiles, a subset of 5,257 genomes were identified which contained no instances of truncated or absent loci (Figure 2.3). To produce a dataset free of untypable loci by discarding loci from the scheme rather than genomes, approximately half of all loci would need to be eliminated (Figure 2.4).

A total of 697 loci were ultimately determined to belong to the *C. jejuni* core genome. One locus identified by ROARY as a core gene, 'GROUP_6337', was manually

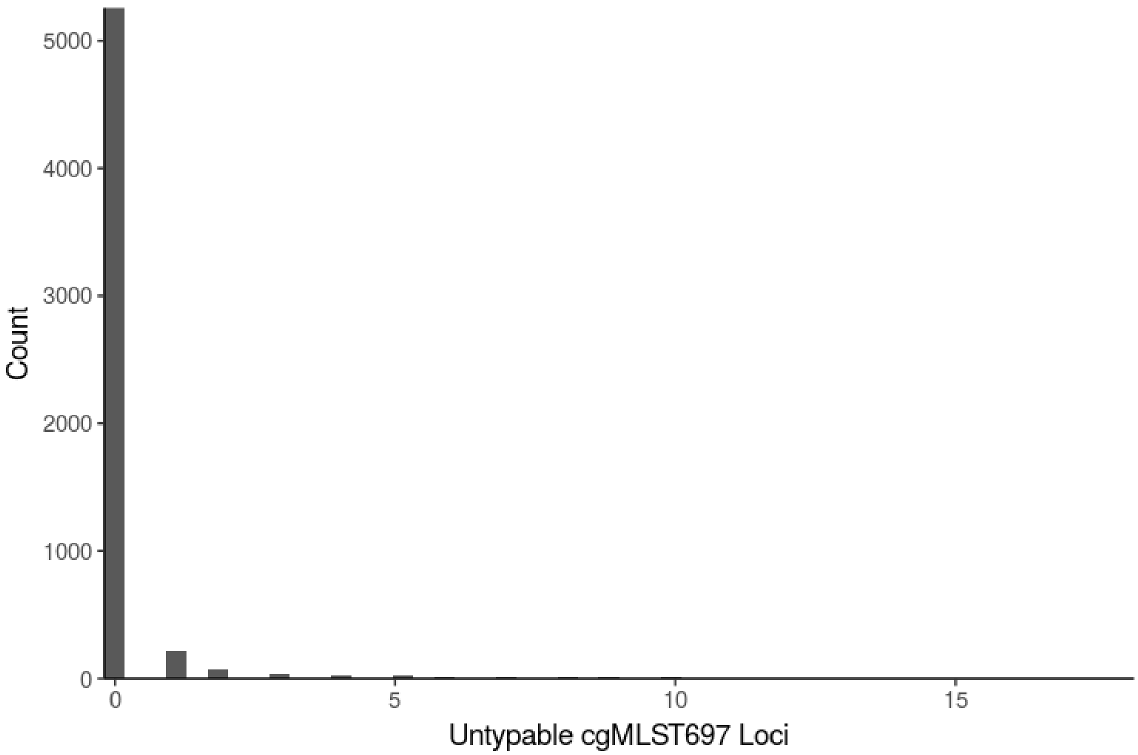

Figure 2.2: A histogram showing the number of untypable cgMLST loci in 5,693 *C. jejuni* draft genomes. The majority (92%) are fully typable at all 697 loci.

removed from the core definition as only partial sequence data could be recovered for 703 genomes.

2.4 Discussion & Conclusions

Developing a robust cgMLST scheme is an important goal for contemporary public health surveillance programs targeting pathogenic microbes. The gene-by-gene approach to microbial genomic epidemiology is highly appropriate to recombinogenic organisms such as *C. jejuni* [117]. The additional discriminatory power relative to classical MLST afforded by cgMLST allows investigators to gain a detailed look at the relationships amongst strains.

The cgMLST scheme described here is more conservative in its design than its peer schemes, such as those described by Cody, *et al.* [93, 94]. These combined *C. jejuni* and *C. coli* cgMLST schemes take differing approaches to the problem of

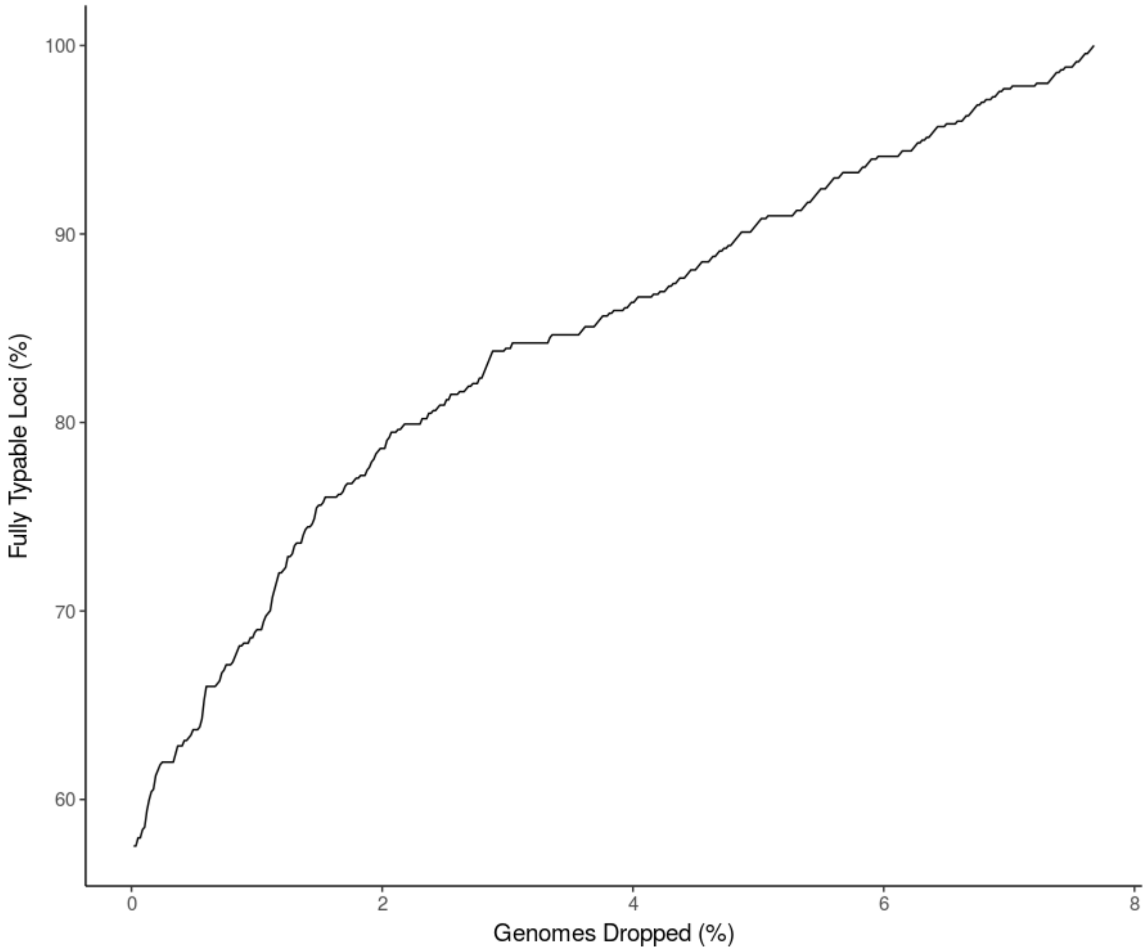

Figure 2.3: The proportion of 5,693 genomes which must be excluded from the analysis to produce 697 fully typable cgMLST loci.

missing data. The earlier scheme begins with a larger set of 1,667 potential loci and then uses an *ad hoc* subset of these comprising those loci which are common to the particular strains under investigation [93]. The later scheme uses a lower threshold for core inclusion, 95% presence, and missing loci were considered to be a match to all other alleles for the purpose of pairwise distance calculations [94]. By doing so, these schemes deviate from the original MLST methodology. In turn, this complicates interpretation and portability of results.

This 697 core locus scheme for *C. jejuni* is a useful testbed for cgMLST development. By minimizing the effect of missing data, it allows for a less biased study of discriminatory power and population partitioning. As will be described in Chapter 3,

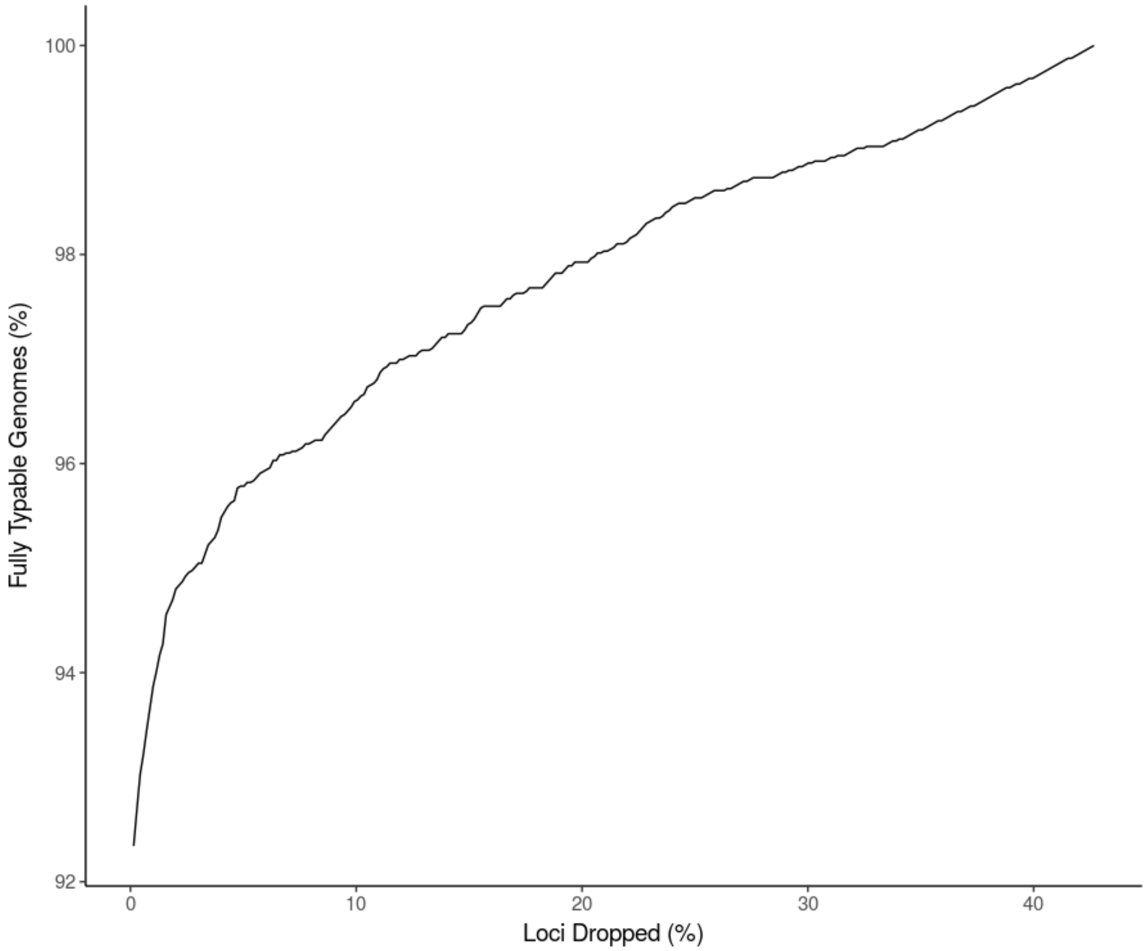

Figure 2.4: The proportion of 697 CGMLST loci which must be excluded from the analysis to produce 5,693 genomes with no missing data.

the number of loci included in a scheme has a direct relationship with both the number of missing loci and the overall discriminatory power of that scheme.

Chapter 3

A Subset of cgMLST Genes Can Recapitulate the Population Structure of the Complete Core Genome

3.1 Introduction

A firm knowledge of population structure is of particular interest to public health investigators when studying pathogenic bacteria. An accurate, high-resolution description of an organism's population structure aids in understanding the transmission dynamics, source attribution, and epidemiology of an organism. Currently, such understanding is generally achieved through the application of molecular methods discussed in Chapter 1.2, particularly those which determine diversity on the basis of the genome sequence. Modern whole genome sequencing (WGS) technologies allow rapid and inexpensive characterization of nearly the complete nucleotide sequence of a given bacterial isolate [126]. The wealth of data already generated and data soon to come create new opportunities for analysing the population structures of a range of bacteria species at a level of detail greater than was previously possible.

Historically, multilocus sequence typing (MLST) was a popular molecular method for *C. jejuni* research [67, 118]. The *C. jejuni* scheme assessed the allelic diversity of seven core genes. A major advantage of the MLST approach is that the allele definitions are readily portable between laboratories. Additionally, each unique allele profile serves as a Sequence Type (ST). An ST must have allele typing data for all loci in the scheme. STs are undefined for profiles with missing or truncated loci [57]. To achieve this portability for STs, having full-length high-quality sequences for all seven target loci is a stringent requirement. Allele collections and ST definitions required manual curation to ensure quality and portability [57, 65].

Combining modern high-throughput WGS techniques and a modern understanding of the bacterial pangenome with the MLST concept leads naturally to the notion of a core genome MLST, or cgMLST scheme, in which most or all of the core genome is used for generating an allele profile [92–94]. However, due to the inherent incompleteness of draft genome assemblies, it may not be possible to recover the full length of all cgMLST loci from WGS data for every isolate. These difficulties arise when a target locus either extends beyond the end of a contig and is truncated, or falls wholly between two contigs and is missing entirely. In these scenarios, the allele present at the affected locus is not directly recoverable without, at minimum, resequencing the isolate.

The two most important attributes of any modern high-throughput typing system in public health use are that it reliably produces stable types, and that those types be sufficiently discriminatory that very similar strains can be distinguished from one another. These attributes are oppositional. As the number of target loci increase, there are more points of comparison, and thus more discriminatory power available to the scheme. However, each additional locus also represents an opportunity for errors to arise.

Given my hypotheses that, a) the observed number of truncated and missing loci scales linearly with the number of loci in the cgMLST scheme, and b) there exists a diminishing marginal utility for the number of target loci included in a cgMLST scheme with respect to discriminatory power, I attempt in this study to determine an optimum number of target genes for a robust cgMLST scheme for *C. jejuni* which balances discriminatory power with typability.

3.2 Methods

3.2.1 Datasets

The dataset comprised 5,693 *C. jejuni* draft genome assemblies, as described in Chapter 2. The assemblies were constructed from raw sequence reads collected from SRA. Sequence assembly and quality control was performed using INNUca [120]. Using the Microbial *In Silico* Typer (MIST), genome assemblies passing quality control had allele calls generated for 697 cgMLST loci [124].

A 'pristine' dataset of 5,257 genomes was created from the original dataset. The pristine dataset was defined as the proper subset of genomes from the original dataset in which no cgMLST loci were truncated or missing. All genomes in the pristine data must have assignable full length alleles for all loci. A single locus which had contig truncations in the majority of genomes was removed from the cgMLST locus set to improve the number of genomes recovered.

3.2.2 Characterization of Missing Data

Truncated and missing data were identified during allele calling. When a MIST-directed blastn alignment of a query cgMLST locus to a subject genome had an expect value of at most 10, the query sequence was not found in its entirety, and the partial query sequence found found at the end of a contig, it was considered to be an instance of a contig truncation. A locus for which no alignment could be found was considered to be missing from the assembly.

For each gene in the 5,693 genome dataset, the number of genomes that were truncated or missing that position was empirically quantified. Because measuring the distribution of missing data in all k-sized permutations of an n-sized set of core genes would run in factorial time, it is necessary to instead estimate the distribution through random sampling. To estimate the prevalence of each type of missing data for a given sample size, genes were randomly selected using the same algorithm and seeds when sampling from the pristine set. This ensures that the same genes will be drawn when comparing rates of absent and truncated loci as when measuring the discriminatory power of those genes.

3.2.3 Monte Carlo Sampling of Gene Subsets

To assess the performance of various subset sizes of the 697 cgMLST genes, a Monte Carlo sampling approach was used to estimate the clustering of *n* genes. Genes were drawn from the 5,257 genome pristine cgMLST typing data. Each *n*-gene subset functions as its own cgMLST scheme. The Monte Carlo simulation was implemented in the R statistical programming language, version 3.3.1 [127]. Pairwise allelic Hamming distances between genomes were calculated with the assistance of R's APE package [66, 128].

For each gene subset size, 10,000 sampling replicates were performed. Each replicate selected *n* genes such that there was no replacement within a replicate. To guarantee reproducibility, for replicate number *i*, the value *i* was used as the seed for pseudo-randomly selecting genes for the subset. The sampling algorithm also ensured that for the same *i*, the selected *n* genes would be a proper subset of any larger value of *n*. For example, if $i = 2$ and $n = 3$, we may select genes *[X, K, C]*. When instead $i = 2$ and $n = 4$, we would then select genes *[X, K, C, D]*.

3.2.4 Cluster Comparison

To compare the gene subsets against the full cgMLST scheme, the cgMLST scheme was clustered using single-linkage clustering at all possible thresholds. This is to say that these reference clusters were defined for each distance *d* such that no member of a given cluster was less than *d* pairwise allele differences from the most closely related member of any other cluster.

Single-linkage hierarchical clusters were also generated for each subset replicate. Reference thresholds are given as the minimum number of pairwise allele differences between the two strains before they agglomerate into the same cluster. Due to the transitive property of single-linkage clustering, it is possible that two strains in the same cluster may have a pairwise distance greater than that of the clustering threshold. However, this produces clusters that are unambiguously distinct from one another. Cluster agreement between gene subsets and the various reference thresholds was calculated using the Adjusted Wallace Coefficient (AWC) [129]. The clusters formed by each subset replicate were compared against the reference thresholds in a pairwise manner.

3.2.5 Locus Partitioning Redundancy

The genome partitioning created by each locus were compared against those of every other locus in order to measure their congruence and redundancy. This was accomplished using the Adjusted Wallace Coefficient [129].

Loci were clustered by their mutual AWC. For example, if two genes A B had $AW_{A \to B}$ and $AW_{B \to A}$ that were both greater than the threshold, they would be placed in the same co-partitioning group. These co-partitioning groups represented collections of genes which partitioned the genome dataset at least as similarly as a given AWC threshold.

3.3 Results

3.3.1 Characterization of Missing Data

The number of missing and truncated loci for a given selection of cgMLST loci has a direct linear relationship to the number of genes selected.

Truncations made up the majority of untypable loci (Table 3.1). Across all subsets and all replicates, the mean and median proportion of missing data stemming from sequencing truncations were 88.9% and 89.6%, respectively. The mean increased slightly with sample size, growing from 85.9% at 7 genes to 89.6% at 650 genes. The median

Table 3.1: Proportion of missing data caused by sequencing truncations for different numbers of selected genes drawn from 697 cgMLST genes over 10,000 replicates.

Genes	Mean	Median	Std. Dev.
7	0.859	1.000	0.216
21	0.869	0.909	0.132
50	0.880	0.900	0.085
100	0.887	0.898	0.057
150	0.890	0.897	0.044
200	0.892	0.897	0.037
250	0.893	0.897	0.031
300	0.894	0.896	0.027
348	0.894	0.896	0.023
400	0.895	0.896	0.020
450	0.895	0.896	0.017
500	0.895	0.895	0.015
550	0.895	0.895	0.012
600	0.895	0.895	0.009
650	0.896	0.895	0.006
All	0.889	0.896	0.074

value exhibited the opposite effect, with 100% of missing data at 7 genes being due to truncations, shrinking to 89.5% at 650 genes (Table 3.2). At locus subset sizes greater than 100, all subsets were affected by missing data (Table 3.3).

3.3.2 Pristine Dataset

Following removal of all genomes containing missing data, the pristine dataset was found to contain 5,257 genomes and 697 genes.

3.3.3 Monte Carlo Simulation of Gene Subsets

For each subset size, and for each replicate, the single-linkage clusters were compared to those of every reference cgMLST threshold. The AWC of subset clusters *versus* reference thresholds quickly approach 1.0, the point at any two strains clustered together by the subset clusters certainly group together at the relevant reference thresh-

Table 3.2: Summary statistics for the observed number of truncated, wholly absent, and total missing data for 10,000 of randomly sampled 697 genes from 5,693 genomes.

	Truncated			Absent			Total		
Genes	Median	Mean	Max.	Median	Mean	Max.	Median	Mean	Max.
7	6.0	11.1	181.0	0.0	1.3	26.0	7.0	12.4	181.0
21	24.0	33.2	235.0	3.0	3.9	29.0	28.0	37.0	237.0
50	67.0	78.6	331.0	8.0	9.2	39.0	77.0	87.8	345.0
100	146.0	157.6	408.0	17.0	18.3	54.0	164.0	175.9	431.0
150	226.0	236.2	530.0	26.0	27.5	64.0	254.0	263.7	562.0
200	307.0	315.2	590.0	35.0	36.6	73.0	344.0	351.8	627.0
250	389.0	393.7	675.0	45.0	45.7	87.0	435.0	439.5	717.0
300	470.0	472.2	737.0	54.0	54.9	93.0	525.0	527.1	810.0
348	548.0	548.3	798.0	64.0	63.8	102.0	613.0	612.0	860.0
400	633.0	630.7	871.0	74.0	73.3	108.0	707.0	704.1	935.0
450	715.0	709.2	933.0	83.0	82.5	114.0	797.0	791.7	1005.0
500	796.0	788.2	978.0	93.0	91.8	120.0	887.0	880.0	1077.0
550	876.0	867.2	1016.0	103.0	101.0	124.0	977.0	968.2	1124.0
600	958.0	946.8	1059.0	112.0	110.2	127.0	1068.0	1057.0	1169.0
650	1038.0	1025.7	1091.0	121.0	119.3	128.0	1156.0	1145.0	1214.0

old, as the reference threshold is relaxed. The 5th percentile of the AWC distribution of the replicates for each subset size were examined. This shows that the in 95% of cases, an MLST-like scheme of that size would produce AWC at least as large *versus* the 697 gene cgMLST scheme (Figure 3.1).

As a measure of efficiency for each gene, the threshold at which the 5th percentile of each subset size achieved an AWC of 1.0 *versus* the complete cgMLST scheme. Increasing the number of genes in a subset greatly increased the discriminatory power when subset sizes were small. At subset sizes of 200 and above, the rate at which gene added discriminatory power to the scheme flattened (Figure 3.2).

3.3.4 Locus Partitioning Redundancy

Groups of loci which congruently partitioned the dataset were identified when their bidirectional AWC exceeded each of the specified range of AWC thresholds (Fig-

Table 3.3: The number of replicates out of 10,000 in which no loci were affected by truncated or missing loci and the number of replicates in which all selected loci were affected. No replicates had all loci or no loci affected for sample sizes greater than 100 genes.

	No Loci Affected			All Loci Affected		
Genes	Truncated	Absent	Total	Truncated	Absent	Total
7	309	5360	190	10	0	198
21	0	1502	0	0	0	1
50	0	90	0	0	0	0
100	0	1	0	0	0	0
150	0	0	0	0	0	0

ure 3.3). The greatest number of co-partitioning groups were found at AWC thresholds of 0.61 to 0.62 (Figure 3.4). Both the number of non-singleton linkage groups and the mean number of members declined toward a AWC cutoff of 1.0, at which all groups are singletons and no gene perfectly replicates the genome partitioning any other.

3.4 Discussion & Conclusions

Choosing the number of loci in a cgMLST scheme involves a balance between two competing factors: resolution and reliability. Including a greater number of loci will improve the ability for a cgMLST scheme to distinguish between two closely-related strains, whilst also adding to the risk of failure due to imperfect whole genome sequence data. As such, any locus included in a cgMLST scheme must contribute enough discriminatory power to justify its inclusion. The increased discriminatory power of each additional locus can be measured by quantifying the degree of redundancy given by bidirectional AWC between the genome partitioning given by the allele distribution of a given pair of loci.

Single-linkage clustering has a long history of use in describing phylogenetic relationships between organisms [130]. An advantage of single-linkage clustering is that it guarantees that an individual strain will never be more closely related to a member

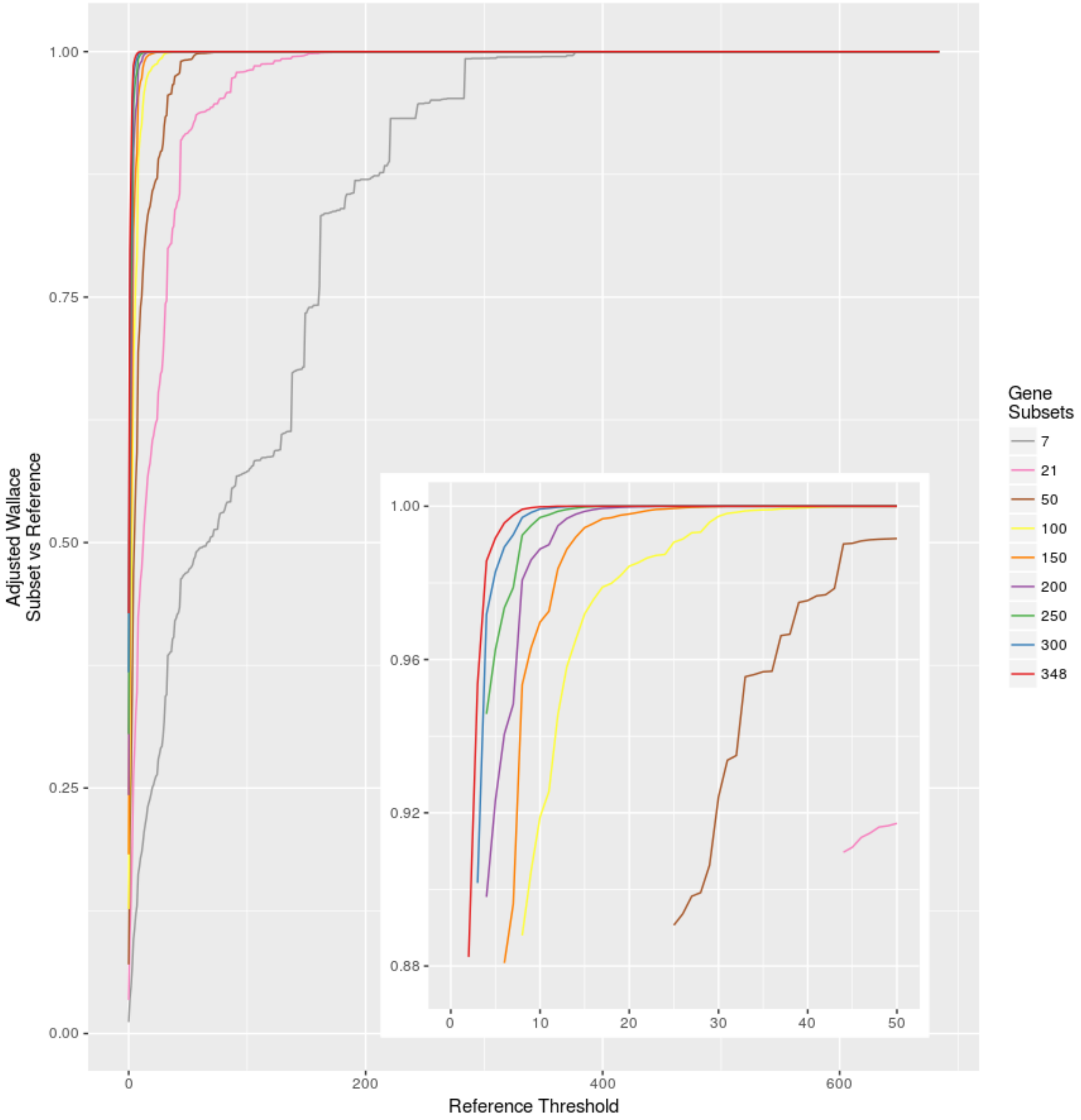

Figure 3.1: *(Main)* The 5th percentile of Adjusted Wallace Coefficient scores of sampled subsets *versus* the complete cgMLST scheme across all clustering thresholds. *(Inset)* A magnification of the main plot showing the region of high subset AWC at stringent cgMLST thresholds.

of another cluster than to the most closely related member of its own cluster. This ensures that clusters are unambiguously distinct from one another.

This study demonstrates three key findings pertinent to the development of a robust cgMLST scheme. The first is that untypable loci are pervasive, even in a dataset consisting of high-quality draft genome assemblies. Even when selecting only 7 loci,

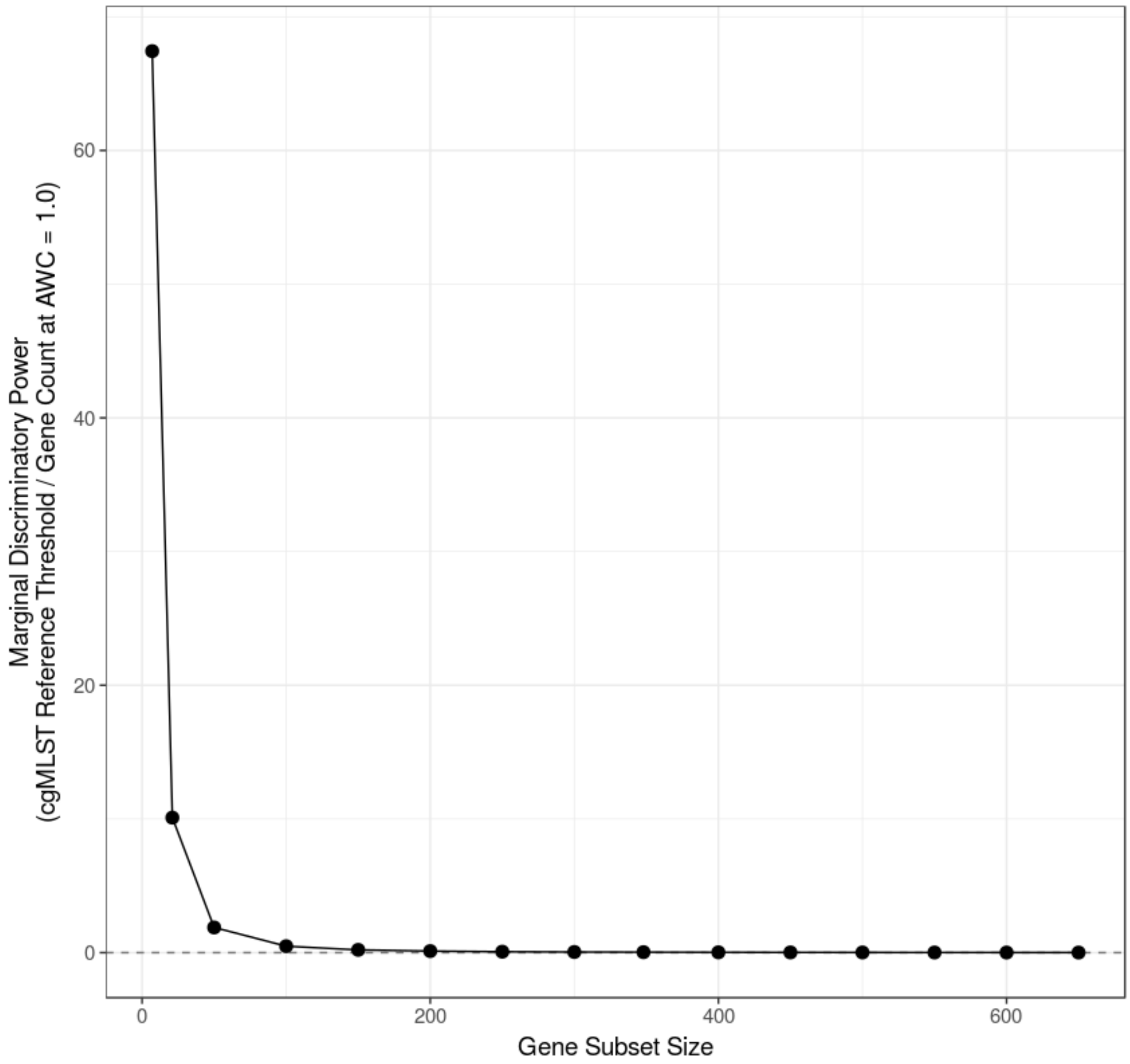

Figure 3.2: The marginal discriminatory power of genes as gene subset size increases. Marginal discriminatory power is given as the number of clustering thresholds per subset gene by which the cgMLST clusters must be relaxed for the subset to achieve an AWC of 1.0 *versus* cgMLST.

just 190 sampling replicates out of 10,000 had complete typing data at all loci for all 5,693 genomes. Scheme sizes of 21 genes and above had no observed replicates with all loci completely typable. The total number of untypable loci, both by truncation and absence, demonstrated a linear relationship to the number of loci selected. The linearity of this relationship stands in contrast to the non-linear relationship between number of loci and discriminatory power.

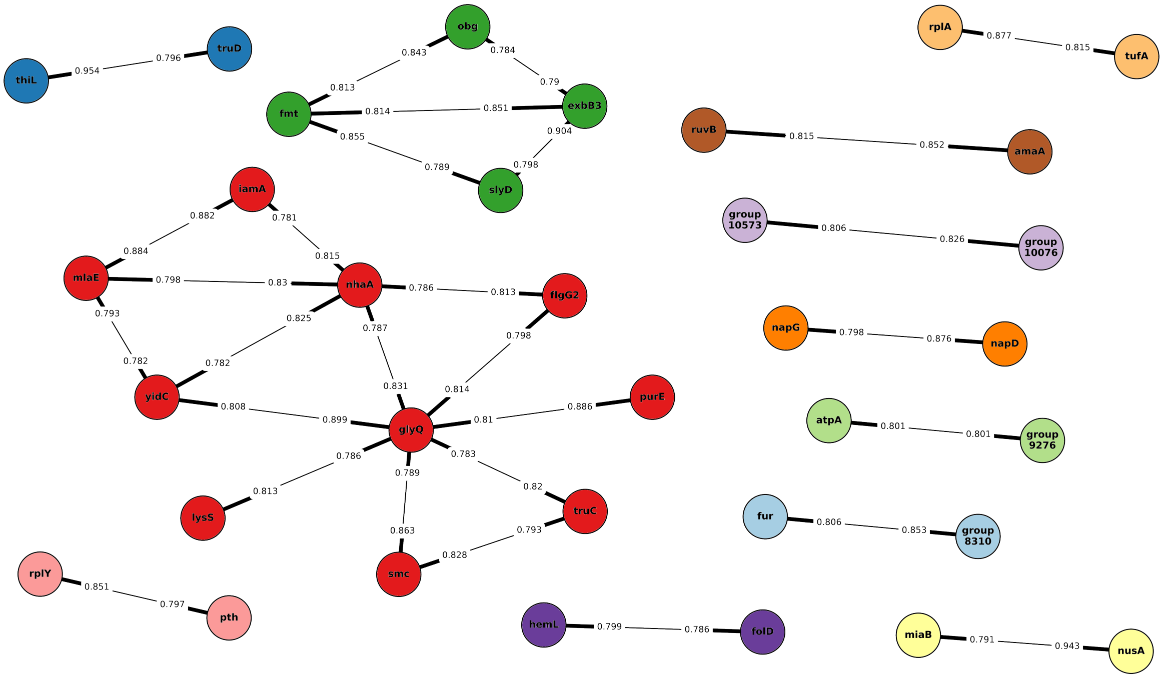

Figure 3.3: A directed graph showing pairs of cgMLST locus genes with an bidirectional Adjusted Wallace Coefficient greater than 0.78, organized into 12 clusters. Genes linked either directly or by the transitive property are assigned the same colour. For a pair of genes A and B, the value given on the edge nearest B represents $AW_{A \to B}$.

The second major finding is that in genome assemblies produced by current DNA sequencing technologies, contig truncation is by far the more common cause of incomplete sequence data *versus* the sequence being wholly missing. This can be interpreted as strong evidence that the loci selected for inclusion in the cgMLST scheme are indeed part of the core genome, and most missing data is due to insufficient read coverage and not due to biological absence.

The third major finding is the relatively small number of loci required to recapitulate the complete locus set at a high clustering threshold. Beyond 200 genes, the amount of additional discriminatory power per gene diminished dramatically. The available evidence suggests that each locus included in a cgMLST-like scheme has diminishing marginal utility with respect to discriminatory power. Additionally, collections of genes exist which have a high bidirectional Adjusted Wallace Coefficient, indicating that these groups of loci partition a diverse genome dataset in a largely

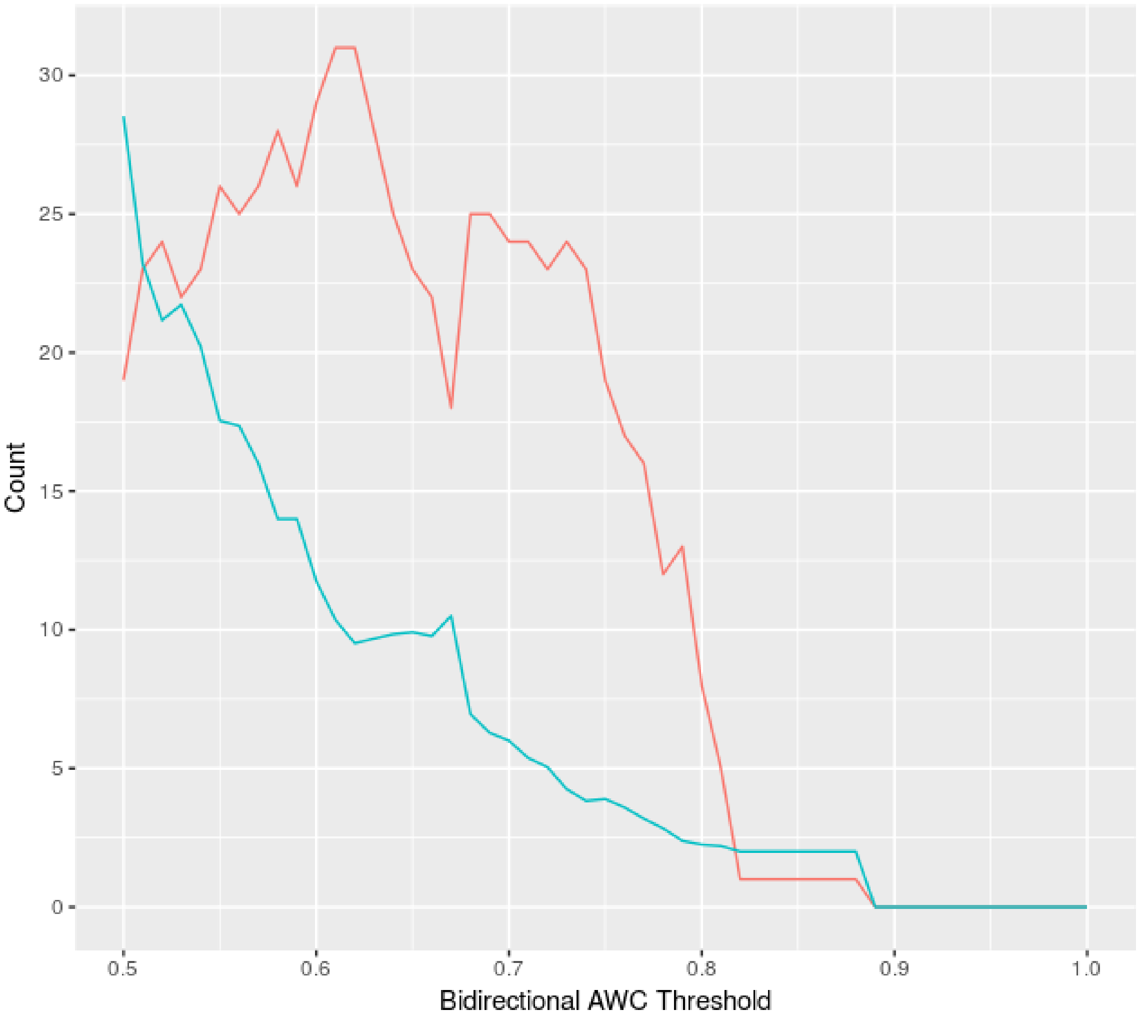

Figure 3.4: *(Red)* The number of groups of genes linked by bidirectional AWC. *(Blue)* The mean number of members of linkage groups.

redundant manner. Linkage disequilibrium between the genes used in this CGMLST would be a plausible explanation for observed partitioning redundancy.

Accurate assessment of these factors demands a large and diverse genome set. Failure to do so can lead to overly optimistic estimates of core genome size. In a recent paper, Cody *et al.* describe a joint CGMLST scheme for *C. jejuni* and *C. coli*. A much smaller number ($n = 2{,}472$) of human clinical isolates isolates from a geographically restricted area comprised the scheme development dataset. This scheme defined 1,343 loci as core genes at a much less stringent threshold of $\geq 95\%$ presence within a dataset of only 2,472 genomes, all of which were isolated in the United Kingdom

[94]. A larger core genome calculated from a smaller, multi-species but epidemiologically homogeneous dataset will necessarily suffer from more missing loci than a more conservative set of loci drawn from a larger, epidemiologically diverse single-species genome collection. It is likely that a CGMLST scheme defined with the parameters of the Cody *Campylobacter* CGMLST scheme will include loci which are not truly core, and instead belong to the accessory genome of the genus.

The fundamental trade-off between discriminatory power and reliability of results should be in the mind of anyone undertaking the task of designing a robust core genome multilocus sequence typing scheme. A greater number of target loci better resolves highly similar strains, but will also introduce uncertainty when loci are inevitably rendered unassignable by incomplete sequence data. Further, a diverse but single-species genome collection is an essential starting point for designing a CGMLST scheme.

Chapter 4

CROWBAR: Bayesian Allele Recovery for Missing Typing Data

4.1 Introduction

Sequence based typing systems such as multilocus sequence typing (MLST) are an important component of modern public health epidemiology and surveillance programmes. MLST and its derived typing systems consider any modification to the nucleotide sequence of a target locus to be a new sequence type. This is equally true for a single nucleotide variant arising from mutation, or the mass import of variable positions following a homologous recombination event. A consequence of this system is that for a novel allele to be described, loci must have complete sequence data available. As such, the loss of even a single base renders the locus untypable for the purposes of MLST [57].

A major problem for current implementations of core genome multilocus sequence typing systems is their inherent susceptibility to data loss by sequencing truncation. As described in Chapter 3, as the number of loci in a scheme increases, the risk of untypable loci increases proportionally. As the size of the dataset expands, data loss becomes a certainty. Because most untyped loci are the result of contig truncation, partial sequence data are often available. Additionally, we found that different CGMLST loci often generated congruent partitions. These facts open the door to the possibility of probabilistically inferring untypable locus calls. In many cases, even when a core genome locus is entirely missing, evidence may exist by which one may infer the identity of the untyped locus.

It is possible to take these lines of evidence — partially recovered sequence data, syntenic relationships between loci, and allele data from the most closely related genome sequence — and use them to inform predictions of the identity of untypable cgMLST alleles.

Here I present crowBAR, a system for probabilistically recovering missing allelic typing data lost due to technical error. This system improves upon MLST-like systems by overcoming their principal drawback at scale: the necessity for complete data at all loci.

4.2 Methods

4.2.1 Fragment Matching

When a locus is untypable due to a sequencing truncation, a partial sequence for that locus is often still recoverable. Though the partial match cannot be used for positive identification of the missing allele, it can be used to assign probabilities to possible identities of the allele. Fragment matching is particularly useful for setting that probability to zero.

When a partial match is returned for a given locus, that match and its reverse complement are queried against all known alleles for that locus. Alleles not containing either form of the query sequence as a substring can be eliminated from subsequent consideration for allele recovery. Alleles that do contain the query substring are assigned a probability based upon their relative abundances within the cgMLST dataset defined in Chapter 2. Matches are only attempted at the beginning and end of each known allele sequence.

4.2.2 Allelic Abundance & Novel Allele Probability

To estimate the probability of a novel allele, we perform a Monte Carlo simulation of allele discovery rates. First, we shuffle a list of allele types for the population. Then, for each element in the list, we test whether the allele present at that element has so far been observed or not. This process is repeated for n iterations, and the proportion of times that a new allele is observed at that element is calculated. The mean value of the last percentile of observations is taken as the probability of a new allele on the next observation.

Allelic abundance in the population is calculated as a simple fraction of the population size. After fragment matching determines which alleles are possible, the abundances are adjusted and probabilities are reallocated by adjusting the denominator to reflect the removal of any alleles which have been excluded by fragment matching, and accounting for the probability of a novel allele.

4.2.2.1 Linkage Disequilibrium

Linkage disequilibrium between loci is exploited to probabilistically identify a missing or incomplete locus. The position of each locus is determined within a reference genome selected by the user. The table of allele calls for all loci is sorted to reflect this order. For each locus, we study a triplet comprising the centre gene and its flanking pair of genes. A contingency table of the alleles of the centre gene versus the alleles of the gene pair of its neighbours is constructed.

Given a particular allele a and v, an N-length vector of allele calls for a particular gene, let v' be a logical vector of the same length as v such that v'_i is given by:

$$v_i^{'} = \begin{cases} 1 & v_i = a \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

We construct the logical vectors l, c, and r in the same manner as v' from the allele vectors of the left, centre, and right genes of the triplet, and using the alleles of the query strain's left flank, the hypothesis allele, and the query strain's right flank, respectively.

Our flanking allele likelihood for our hypothesis allele, h, is then given by:

$$P(flank|h) = \frac{\sum_{i=0}^{N} l \wedge c \wedge r}{\sum_{i=0}^{N} l \wedge r} \tag{4.2}$$

4.2.2.2 Closest Neighbour

Because the closest relative of a strain necessarily shares more alleles with the query strain than is average, an additional source of evidence as to the identity of the missing locus is that of its nearest neighbour. If multiple strains are equidistant to the query strain, all observations are considered equally.

For each hypothesized allele, h, if h is observed amongst neighbouring strains within the dataset, we determine its probability of being present in the query strain as:

$$D = (1 - abundance_h{}^N) * d_{neighbour} \tag{4.3}$$

$$D' = (1 - d_{neighbour}) * abundance_h \tag{4.4}$$

$$neighbour = \begin{cases} D & N > 0 \\ D' & otherwise \end{cases} \tag{4.5}$$

where N is the number of observations of h, and $d_{neighbour}$ is the distance between the query strain and its closet neighbour expressed as the number of differing loci divided by the total number of loci.

4.2.3 Allele Recovery

The various likelihoods described above are combined using Bayes' Theorem to return a probability for each hypothesis allele [131]. Each allele is tested as a hypothesis independently. The relative abundances of each allele, updated to reflect the outcome of fragment matching, are used as the prior probabilities of each hypothesis.

$$P(E|H) = neighbour * flank \tag{4.6}$$

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{4.7}$$

where *neighbour* and $flank$ are found using Equations (4.5) and (4.2), and $P(E)$ is the sum of the likelihoods given by Equation (4.6) for all hypothesis alleles.

4.2.4 Implementation

The CROWBAR system is implemented in Python 3, and makes use of the NumPy and Pandas libraries to assist with numerical calculations and handling tabular data [132–134].

The first steps taken by CROWBAR are to ensure that the table of CGMLST calls are placed in the same relative order as they are found in a user-selected reference genome. This is accomplished by using BLASTN to locate a representative of each gene in the reference and reordering the table by the genome locations of the alignments [100].

Next, a Hamming distance matrix is calculated [66]. A pre-calculated matrix may also be provided. From this matrix, the closest relatives of each strain are determined. For any given pairwise comparison, loci in which a truncation or missing locus is present are excluded from that comparison. This distance matrix is used to determine the closest relative of the strain under examination, and the degree of similarity between the two strains.

Allele abundances are determined, and then adjusted to reflect the probability of a previously undescribed allele being observed. If partial sequence data are available for a locus, it can be used to eliminate alleles which are not possible because they do not contain the observed partial sequence. After alleles have been removed from consideration, the relative abundance of alleles is calculated, including the probability of discovering a novel allele.

Flank linkage likelihoods are determined as discussed above. In the event that the hypothesis allele is never observed with the query strain's flanking gene alleles, or if the flanking gene alleles are also untypable, the novel allele probability is returned instead.

For each truncated or absent allele, the likelihoods of each line of evidence are calculated as described above. Bayes' Theorem combines these with the prior for each hypothesis allele — the allele's abundance adjusted for fragment matching — and returns their probabilities given the evidence.

The source code for CROWBAR is freely available at:

`https://github.com/dorbarker/crowbar.git`

4.2.5 Validation

To validate CROWBAR and measure its success rate, we designed a second script to manage creation and checking artificially truncated and missing data. A table of allele calls for 697 loci in 5257 *C. jejuni* genomes with complete allele typing data at all loci was used to generate test data for this experiment. To reorder the calls table, we used *C. jejuni* NCTC11168, a common reference genome for the species [109]. Loci were randomly truncated or rendered absent with probabilities of 0.3% and 0.035%, respectively. These probabilities were selected to reflect the empirical rates of truncation observed in Chapter 3. Truncations were created such that fragments were never less than 50 bp in length. For reproducibility, the algorithm which controls error introduction proceeds deterministically from a seed value set by the user. Estimating the probability of a novel allele ran for 1,000 Monte Carlo iterations. For each synthetic error, crowBAR was used to recover the underlying allele and returns the most probable candidate. Because the errors are synthetic, their true identities are known and can be compared against the recovered allele.

4.3 Results

Table 4.1: Recovery rates using crowBAR for ten replicates of synthetic errors in 5,257 genomes and 697 genes. Truncations and absent loci were randomly applied at rates of 0.03% and 0.0035%. False Novel indicates the percentage of errors which were not successfully recovered that were determined to be a novel allele by crowBAR.

				False	Success	False
Replicate	Truncations	Absent	Successes	Novel	Rate	Novel Rate
1	1122	132	1147	49	91.47	45.79
2	1081	134	1118	44	92.02	45.36
3	1113	133	1139	42	91.41	39.25
4	1117	147	1178	38	93.20	44.19
5	1077	119	1109	33	92.73	37.93
6	1079	135	1141	32	93.99	43.84
7	1105	140	1161	35	93.25	41.67
8	1096	133	1140	34	92.76	38.20
9	1088	123	1118	39	92.32	41.94
10	1127	111	1160	41	93.70	52.56
Overall	11005	1307	11411	387	92.68	42.95

Table 4.1 shows the performance of crowBAR over ten replicates of randomly applied synthetic truncations and missing loci to the pristine dataset described in Chapter 3. Success rate gives the proportion of recovery attempts which successfully induced the identity of the underlying allele. On average, 92.68% of attempts were successful. The worst performing replicate had a success rate of 91.41%. Amongst the cases in which the allelic identity of the locus was not correctly ascertained, 42.9% of these failures resulted from the spurious reporting of a novel allele.

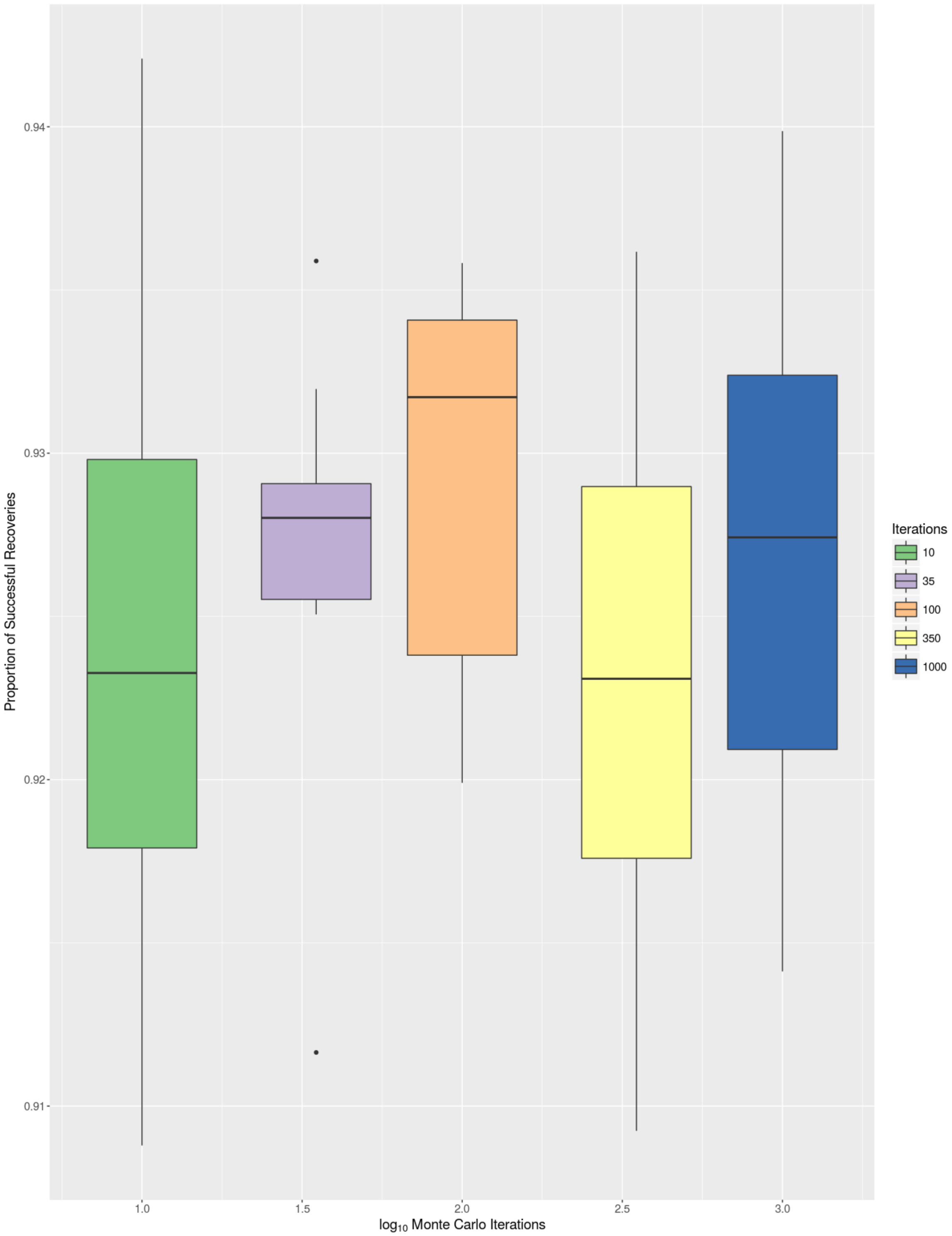

Figure 4.1: The apparent lack of effect of increasing the number of Monte Carlo iterations in estimating the probability that a untypable locus is an undescribed allele.

I observed no effect from varying the number of Monte Carlo iterations during the estimation of novel allele probability. Figure 4.1 shows fully overlapping box plots of the distribution of success rates for 10 replicates each of 10, 35, 100, 350, and 1,000 Monte Carlo iterations.

The number of genomes available to CROWBAR determines the effectiveness of the algorithm. Figure 4.2 shows the results of a Monte Carlo simulation of 1,000 iterations sampling a number of genomes in the interval [100, 5,257). The number of genomes, g, appears to be related to the recovery success rate S by $S = log(g)$.

4.4 Discussion & Conclusions

While calculating allele abundance, CROWBAR repeatedly shuffles the list of alleles for that locus. The intent of this process is to provide an unbiased estimate for the probability of the missing locus being a novel allele. By using the mean allele discovery rate of the last percentile of observations, this approximates the probability that the next observation will be a previously unobserved allele. Surprisingly, the number of shuffling steps does not appear to be important to the accuracy of the results given this experimental dataset. However, several factors necessitate the inclusion of this step. Though we can estimate the total number of alleles for a locus using nonparametric estimators such as the Chao 1 Estimator, even in a closed population, loci are mutable and novel alleles can arise at any time [135]. Thus, the probability of an untyped locus possessing a novel allele must never be zero. As sequencing efforts continue and allelic diversity is more fully explored, the rate of allele discovery may fall to a point such that without a shuffling step, the probability of a novel allele may incorrectly be set to zero.

In the 7.32% of cases where CROWBAR failed to recover the allele, nearly half were falsely predicted to be novel alleles. This raises a quandary. As a control, this particular experiment uses synthetic errors introduced to perfect data. Because the

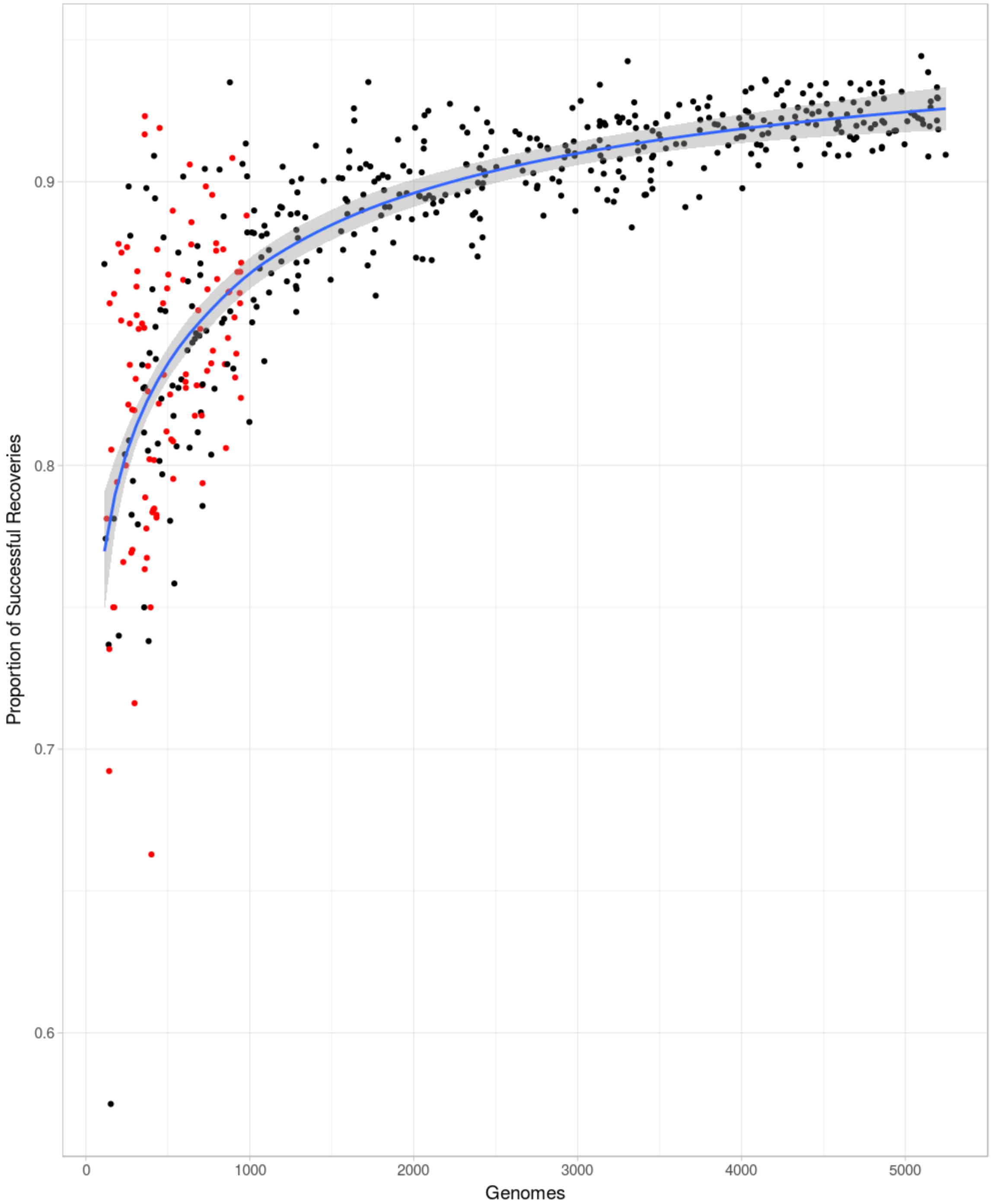

Figure 4.2: The relationship between the number of genomes analyzed and its effect on the success rate of CROWBAR. The blue line indicates the fit line for $S = log(g)$ and the grey shaded area is the 99% confidence interval for the standard error of the fit.

identity of every locus in every strain is known, no novel alleles exist to be found. However, failure to consider the possibility of a novel allele existing at a untyped locus in a real world dataset would be irresponsible. The probability of a novel allele is always present. However, the frequency with which missing loci are falsely reported to have a novel allele is an obvious target for future improvements to the algorithm.

Traditionally, allelic typing systems like MLST and its derivatives require complete typing data at all loci. Though a system like CROWBAR cannot truly replace high quality whole genome sequence data, it can be used to repair errors in typing data and avoid having to discard otherwise-useful genomes. Completing the set of typing calls allows for a nomenclature to be applied to allele profiles, which is of benefit for inter-laboratory communication and long-term monitoring of strains of interest.

As a fundamental aspect of its operation, CROWBAR implicitly assumes that there is a locus to recover. This experiment uses genes which previous work strongly suggests are core genes. Loci which are truncated or missing are presumed to be so due to technical rather than biological reasons. If CROWBAR is given accessory rather than core genes, it will return a spurious result.

Core genome multilocus sequence typing systems are becoming increasingly prevalent in public health surveillance programmes and microbial source tracking. Though modern sequence platforms are impressive, they cannot be relied upon to generate perfect WGS data, even if current CGMLST doctrine demands it. Though CROWBAR already recovers missing or trucated loci with greater than 90% accuracy, additional genome data will further improve the accuracy of the statistical model used to recover these loci. Continuing global whole genome sequencing efforts will be a ready source of this data. Additional finesse to the model may additionally improve results by reducing over-estimation of novel allele discovery rates. By building a simple but robust statistical model, CROWBAR offers an accurate and reproducible system for recovering loci lost or damaged by sequencing errors.

Chapter 5

Conclusions

As genome sequencing is more widely adopted for characterization of bacterial pathogens, cgMLST has increasingly been put forward as the preferred method for the long term tracking of strains of interest. The majority of historical molecular typing methods used to infer relationships between microbial strains were developed before the advent of inexpensive and reliable DNA sequencing technology, and are described in Chapter 1. These methods had been designed to exploit differences between strains in their macromolecular structure. Detection of these differences typically relied upon susceptibility stressors, electrophoretic mobility, chemical or immunological reactivity, or PCR amplification. All of these are ultimately abstractions of the underlying variation between nucleotide sequences. Classical MLST was developed to type the DNA of conserved core genes directly whilst also controlling for distortions caused by homologous recombination events. With the advent of high-throughput whole genome sequencing, cgMLST emerged as a natural extension of the MLST concept.

As the number of draft genome sequences available in public repositories and private collections continues to increase, so too does the potential utility of a cgMLST scheme. However, as the work presented in Chapter 3 describes, the severity of the problem posed by absent or truncated loci is proportional to the number of loci incorporated into a given scheme. Classical MLST required that there be complete typing data at all loci before a profile could be assigned a Sequence Type in the nomenclature system. Any prototype cgMLST should therefore attempt to minimize the number of untypable loci to ensure both the reliability of the data and the assignment of the profile to an unambiguous nomenclature which represents subpopulations of closely related genomes which maintain a nomenclature designation stably over time.

The work presented in this thesis tackled this in three ways. Chapter 2 describes a methodology for conservatively designing cgMLST schemes. This hinges upon selecting only genes which have higher rates of carriage than competing cgMLST definitions, *e.g.*, 99.9% *versus* 95% presence. In doing so, we can be more confident that all loci included in the scheme are core genes and do not belong to the accessory genome. Also, by eliminating genes with empirically greater than average rates of sequencing truncations, we can improve confidence that all loci will be typable. Chapter 3 describes using subsets of this cgMLST scheme to produce allele profiles which are concordant with the superset. This work involved identifying groups of genes which partition the dataset such that they produce high bidirectional Adjusted Wallace Coefficients. By identifying groups of loci that partition the dataset congruently, genes which can be dropped from the scheme while minimally impacting discriminatory power can be identified. Doing so can further optimize the cgMLST scheme by reducing the probability of a future sequencing truncation occurring within the scheme's selected loci. Defining highly conservative subsets which sacrifice the least discriminatory power in exchange for the greatest reduction in missing loci achieves a desirable attribute in a typing system. Finally, in Chapter 4, I present a system for inferring the identities of cgMLST loci rendered untypable due to technical, rather than biological, reasons. This system is implemented as a tool, crowBAR, which draws its predictions from partial sequence matches to known alleles, to allelic profiles of closely related strains, and from patterns of gene linkage disequilibrium. In combination, these factors are highly effective at predicting known alleles. It is also capable of identifying cases where the locus is likely to be a previously unknown allele, although at present crowBAR currently overestimates the likelihood of this scenario.

Improving the estimation of the likelihood of novel alleles is a promising avenue for future work in improving the model.

Taken together, this research can be used to develop and deploy robust cgMLST systems. This thesis provides rules and good practices to use these schemes in support of public health surveillance programmes. This work represents an important advancement in cgMLST design as we enter the genomic era.

❖

Bibliography

1. Cravioto, A., Lanata, C. F., Lantagne, D. S. & Nair, G. B. *Final report of the independent panel of experts on the cholera outbreak in Haiti* 2011.

2. Hendriksen, R. S. *et al.* Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio* **2**, e00157–11 (2011).

3. Katz, L. S. *et al.* Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* **4**, e00398–13 (2013).

4. Frerichs, R. R., Keim, P. S., Barrais, R. & Piarroux, R. Nepalese origin of cholera epidemic in Haiti. *Clinical Microbiology and Infection* **18**, 158–163 (2012).

5. Murray, C. J. L. *et al.* Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* **380**, 2197–2223 (2013).

6. Sabat, A. J. *et al.* Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveillance* **18**, 20380 (2013).

7. Maslow, J. N., Mulligan, M. E. & Arbeit, R. D. Molecular epidemiology: application of contemporary techniques to the typing of microorganisms. *Clinical Infectious Diseases* **17**, 153–162 (1993).

8. Van Belkum, A. *et al.* Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clinical Microbiology and Infection* **13**, 1–46 (2007).

9. Eberle, K. & Kiess, A. Phenotypic and genotypic methods for typing *Campylobacter jejuni* and *Campylobacter coli* in poultry. *Poultry Science* **91**, 255–264 (2012).

10. Maslow, J. N. *et al.* Relationship between indole production and differentiation of *Klebsiella* species: indole-positive and-negative isolates of *Klebsiella* determined to be clonal. *Journal of Clinical Microbiology* **31**, 2000–2003 (1993).

11. Skirrow, M. & Benjamin, J. Differentiation of enteropathogenic *Campylobacter. Journal of Clinical Pathology* **33**, 1122 (1980).

12. Totten, P. A. *et al.* Prevalence and characterization of hippurate-negative *Campylobacter jejuni* in King County, Washington. *Journal of Clinical Microbiology* **25**, 1747–1752 (1987).

13. Varga, C. *et al.* Comparison of antimicrobial resistance in generic *Escherichia coli* and *Salmonella* spp. cultured from identical fecal samples in finishing swine. *Canadian Journal of Veterinary Research* **72**, 181–187 (2008).

14. Deckert, A. *et al.* Canadian Integrated Program for Antimicrobial Resistance Surveillance (CIPARS) farm program: results from finisher pig surveillance. *Zoonoses and Public Health* **57**, 71–84 (2010).

15. Mulvey, M. R. *et al.* Emergence of multidrug-resistant *Salmonella enterica* serotype 4,[5], 12: i:-involving human cases in Canada: results from the Canadian Integrated Program on Antimicrobial Resistance Surveillance (CIPARS), 2003–10. *Journal of Antimicrobial Chemotherapy,* dkt149 (2013).

16. Craigie, J. & Yen, C. H. The demonstration of types of *B. typhosus* by means of preparations of Type II Vi phage: I. Principles and Technique. *Canadian Public Health Journal* **29**, 448–463 (1938).

17. Cotter, P. D., Ross, R. P. & Hill, C. Bacteriocins: a viable alternative to antibiotics? *Nature Reviews Microbiology* **11**, 95–105 (2013).

18. Abbott, J. & Shannon, R. A method for typing *Shigella sonnei*, using colicine production as a marker. *Journal of Clinical Pathology* **11**, 71–77 (1958).

19. Sell, T. L., Schaberg, D. R. & Fekety, F. R. Bacteriophage and bacteriocin typing scheme for *Clostridium difficile. Journal of Clinical Microbiology* **17**, 1148–1152 (1983).

20. Traub, W., Raymond, E. & Startsman, T. Bacteriocin (marcescin) typing of clinical isolates of *Serratia marcescens. Applied Microbiology* **21**, 837–840 (1971).

21. Lancefield, R. C. A serological differentiation of human and other groups of hemolytic streptococci. *The Journal of experimental medicine* **57**, 571–595 (1933).

22. Penner, J., Hennessy, J. & Congi, R. Serotyping of *Campylobacter jejuni* and *Campylobacter coli* on the basis of thermostable antigens. *European Journal of Clinical Microbiology* **2**, 378–383 (1983).

23. Linton, D., Karlyshev, A. V. & Wren, B. W. Deciphering *Campylobacter jejuni* cell surface interactions from the genome sequence. *Current Opinion in Microbiology* **4**, 35–40 (2001).

24. *Salmonella* Subcommittee of the Nomenclature Committee of the International Society for Microbiology. The genus *Salmonella lignieres*, 1900. *The Journal of Hygiene* **34**, 333 (1934).

25. Ørskov, F. & Ørskov, I. Serotyping of *Escherichia coli. Methods in Microbiology* **14**, 43–112 (1984).

26. Doumith, M., Buchrieser, C., Glaser, P., Jacquet, C. & Martin, P. Differentiation of the major *Listeria monocytogenes* serovars by multiplex PCR. *Journal of Clinical Microbiology* **42**, 3819–3822 (2004).

27. Facklam, R. *et al. emm* typing and validation of provisional M types for group A streptococci. *Emerging Infectious Diseases* **5**, 247 (1999).

28. Murphy, M. *et al.* Development and application of Multiple-Locus Variable number of tandem repeat Analysis (MLVA) to subtype a collection of *Listeria monocytogenes. International Journal of Food Microbiology* **115**, 187–194 (2007).

29. Van Belkum, A. Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA). *Pathogens and Disease* **49**, 22–27 (2007).

30. Sabat, A. *et al.* New method for typing *Staphylococcus aureus* strains: multiple-locus variable-number tandem repeat analysis of polymorphism and genetic relationships of clinical isolates. *Journal of Clinical Microbiology* **41**, 1801–1804 (2003).

31. Spurgiesz, R. S. *et al.* Molecular typing of *Mycobacterium tuberculosis* by using nine novel variable-number tandem repeats across the Beijing family and low-copy-number IS6110 isolates. *Journal of Clinical Microbiology* **41**, 4224–4230 (2003).

32. Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J. A. & Tingey, S. V. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* **18**, 6531–6535 (1990).

33. Welsh, J. & McClelland, M. Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research* **18**, 7213–7218 (1990).

34. Wang, G., Whittam, T. S., Berg, C. M. & Berg, D. E. RAPD (arbitrary primer) PCR is more sensitive than multilocus enzyme electrophoresis for distinguishing related bacterial strains. *Nucleic Acids Research* **21**, 5930–5933 (1993).

35. Cornelius, A. J., Gilpin, B., Carter, P., Nicol, C. & On, S. L. Comparison of PCR binary typing (P-BIT), a new approach to epidemiological subtyping of *Campylobacter jejuni*, with serotyping, pulsed-field gel electrophoresis, and multilocus sequence typing methods. *Applied and Environmental Microbiology* **76**, 1533–1544 (2010).

36. Cornelius, A. J. *et al.* Same-day subtyping of *Campylobacter jejuni* and *C. coli* isolates by use of multiplex ligation-dependent probe amplification–binary typing. *Journal of Clinical Microbiology* **52**, 3345–3350 (2014).

37. Taboada, E. N. *et al.* Development and validation of a comparative genomic fingerprinting method for high-resolution genotyping of *Campylobacter jejuni. Journal of Clinical Microbiology* **50**, 788–797 (2012).

38. Clark, C. G. *et al.* Comparison of molecular typing methods useful for detecting clusters of *Campylobacter jejuni* and *C. coli* isolates through routine surveillance. *Journal of Clinical Microbiology* **50**, 798–809 (2012).

39. Laing, C. *et al.* Rapid determination of *Escherichia coli* O157: H7 lineage types and molecular subtypes by using comparative genomic fingerprinting. *Applied and Environmental Microbiology* **74**, 6606–6615 (2008).

40. Webb, A. L., Kruczkiewicz, P., Selinger, L. B., Inglis, G. D. & Taboada, E. N. Development of a comparative genomic fingerprinting assay for rapid and high resolution genotyping of *Arcobacter butzleri. BMC Microbiology* **15**, 1 (2015).

41. Schwartz, D. C. & Cantor, C. R. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**, 6 (1984).

42. Swaminathan, B., Barrett, T. J., Hunter, S. B., Tauxe, R. V. & Force, C. P. T. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging Infectious Diseases* **7**, 382 (2001).

43. Swaminathan, B. *et al.* Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. *Foodbourne Pathogens & Disease* **3**, 36–50 (2006).

44. Clark, C. G. *et al.* Characterization of waterborne outbreak–associated *Campylobacter jejuni*, Walkerton, Ontario. *Emerging Infectious Diseases* **9**, 1232 (2003).

45. Selander, R. K. *et al.* Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Applied and Environmental Microbiology* **51**, 873 (1986).

46. Lewontin, R. C. & Hubby, J. L. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics* **54**, 595 (1966).

47. Harris, H. Enzyme polymorphisms in man. *Proceedings of the Royal Society of London* **164**, 298–310 (1966).

48. Caugant, D. *et al.* Genetic diversity and relationships among strains of *Escherichia coli* in the intestine and those causing urinary tract infections, 203–227 (1983).

49. Ochman, H. & Selander, R. K. Evidence for clonal population structure in *Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America* **81**, 198 (1984).

50. Achtman, M. & Pluschke, G. Clonal analysis of descent and virulence among selected *Escherichia coli. Annual Reviews in Microbiology* **40**, 185–210 (1986).

51. Whittam, T. S. & Wilson, R. A. Genetic relationships among pathogenic *Escherichia coli* of serogroup O157. *Infection and Immunity* **56**, 2467–2473 (1988).

52. Whittam, T. S., Ochman, H. & Selander, R. K. Geographic components of linkage disequilibrium in natural populations of *Escherichia coli. Molecular Biology and Evolution* **1**, 67–83 (1983).

53. Go, M. F., Kapur, V., Graham, D. Y. & Musser, J. M. Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. *Journal of Bacteriology* **178**, 3934–3938 (1996).

54. Cochrane, B. J. & Richmond, R. C. Studies of esterase 6 in *Drosophila melanogaster.* I. The genetics of a posttranslational modification. *Biochemical genetics* **17**, 167–183 (1979).

55. Johnson, G., Finnerty, V. & Hartl, D. Post-translational modification of xanthine dehydrogenase in a natural population of *Drosophila melanogaster. Genetics* **98**, 817–831 (1981).

56. Tenover, F. C. *et al.* Comparison of traditional and molecular methods of typing isolates of *Staphylococcus aureus. Journal of Clinical Microbiology* **32**, 407–415 (1994).

57. Maiden, M. C. *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences* **95**, 3140–3145 (1998).

58. Meinersmann, R., Helsel, L., Fields, P. & Hiett, K. Discrimination of *Campylobacter jejuni* isolates by *fla* gene sequencing. *Journal of Clinical Microbiology* **35**, 2810–2814 (1997).

59. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463–5467 (1977).

60. Clark, C. G. *et al.* Phylogenetic relationships of *Campylobacter jejuni* based on *porA* sequences. *Canadian Journal of Microbiology* **53**, 27–38 (2007).

61. Cody, A. J., Maiden, M. J. & Dingle, K. E. Genetic diversity and stability of the *porA* allele as a genetic marker in human *Campylobacter* infection. *Microbiology* **155**, 4145–4154 (2009).

62. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences* **74**, 5088–5090 (1977).

63. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72**, 5069–5072 (2006).

64. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal* **6**, 610–618 (2012).

65. Jolley, K. A. & Maiden, M. C. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 1 (2010).

66. Hamming, R. W. Error detecting and error correcting codes. *Bell System Technical Journal* **29**, 147–160 (1950).

67. Maiden, M. C. *et al.* MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology* **11**, 728–736 (2013).

68. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43**, e15–e15 (2014).

69. Foley, S. L. *et al.* Comparison of subtyping methods for differentiating *Salmonella enterica* serovar Typhimurium isolates obtained from food animal sources. *Journal of Clinical Microbiology* **44**, 3569–3577 (2006).

70. Mutschall, S. K. *Quick guide to comparative genomic fingerprinting* (2015).

71. Dhiman, N., Hall, L., Wohlfiel, S. L., Buckwalter, S. P. & Wengenack, N. L. Performance and cost analysis of matrix-assisted laser desorption ionization–time of flight mass spectrometry for routine identification of yeast. *Journal of Clinical Microbiology* **49**, 1614–1616 (2011).

72. Vanhee, L., Symoens, F., Jacobsen, M., Nelis, H. & Coenye, T. Comparison of multiple typing methods for *Aspergillus fumigatus*. *Clinical Microbiology and Infection* **15**, 643–650 (2009).

73. Saghrouni, F., Ben Abdeljelil, J., Boukadida, J. & Ben Said, M. Molecular methods for strain typing of *Candida albicans*: a review. *Journal of Applied Microbiology* **114**, 1559–1574 (2013).

74. O'Farrell, B., Haase, J. K., Velayudhan, V., Murphy, R. A. & Achtman, M. Transforming microbial genotyping: a robotic pipeline for genotyping bacterial strains. *PLoS One* **7**, e48022 (2012).

75. Olsen, J. S. *et al.* Evaluation of a highly discriminating multiplex multi-locus variable-number of tandem-repeats (MLVA) analysis for *Vibrio cholerae. Journal of Microbiological Methods* **78**, 271–285 (2009).

76. Graves, L. M. & Swaminathan, B. PulseNet standardized protocol for subtyping *Listeria monocytogenes* by macrorestriction and pulsed-field gel electrophoresis. *International Journal of Food Microbiology* **65**, 55–62 (2001).

77. Ribot, E. M., Fitzgerald, C., Kubota, K., Swaminathan, B. & Barrett, T. J. Rapid pulsed-field gel electrophoresis protocol for subtyping of *Campylobacter jejuni. Journal of Clinical Microbiology* **39**, 1889–1894 (2001).

78. Bannerman, T. L., Hancock, G. A., Tenover, F. C. & Miller, J. M. Pulsed-field gel electrophoresis as a replacement for bacteriophage typing of *Staphylococcus aureus. Journal of Clinical Microbiology* **33**, 551–555 (1995).

79. Frost, J., Kramer, J. & Gillanders, S. Phage typing of *Campylobacter jejuni* and *Campylobacter coli* and its use as an adjunct to serotyping. *Epidemiology & Infection* **123**, 47–55 (1999).

80. Yoshida, C. *et al.* Evaluation of molecular methods for identification of *Salmonella* serovars. *Journal of Clinical Microbiology* **54**, 1992–1998 (2016).

81. Alakomi, H.-L. & Saarela, M. *Salmonella* importance and current status of detection and surveillance methods. *Quality Assurance and Safety of Crops & Foods* **1**, 142–152 (2009).

82. Dworzanski, J. P. & Snyder, A. P. Classification and identification of bacteria using mass spectrometry-based proteomics. *Expert Review of Proteomics* **2**, 863–878 (2005).

83. Dieckmann, R., Graeber, I., Kaesler, I., Szewzyk, U. & Von Döhren, H. Rapid screening and dereplication of bacterial isolates from marine sponges of the Sula Ridge by Intact-Cell-MALDI-TOF mass spectrometry (ICM-MS). *Applied Microbiology and Biotechnology* **67**, 539–548 (2005).

84. Seng, P. *et al.* MALDI-TOF-mass spectrometry applications in clinical microbiology. *Future Microbiology* **5**, 1733–1754 (2010).

85. Taboada, E. N., Graham, M. R., Carriço, J. A. & Van Domselaar, G. Food safety in the age of next generation sequencing, bioinformatics, and open data access. *Frontiers in Microbiology* **8**, 909 (2017).

86. Machado, M. P., Ribeiro-Gonçalves, B., Silva, M., Ramirez, M. & Carriço, J. A. Epidemiological surveillance and typing methods to track antibiotic resistant strains using high throughput sequencing. *Antibiotics,* 331–356 (2017).

87. Lynch, T., Petkau, A., Knox, N., Graham, M. & Van Domselaar, G. A primer on infectious disease bacterial genomics. *Clinical Microbiology Reviews* **29**, 881–913 (2016).

88. Goswitz, J. J., Willard, K. E., Fasching, C. E. & Peterson, L. R. Detection of *gyrA* gene mutations associated with ciprofloxacin resistance in methicillin-resistant *Staphylococcus aureus*: analysis by polymerase chain reaction and automated direct DNA sequencing. *Antimicrobial Agents and Chemotherapy* **36**, 1166–1169 (1992).

89. Petkau, A. *et al.* SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microbial Genomics* **3** (2017).

90. Bekal, S. *et al.* Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent *Salmonella enterica* serovar Heidelberg clone in the context of outbreak investigations. *Journal of Clinical Microbiology* **54**, 289–295 (2016).

91. Jolley, K. A. *et al.* Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* **158**, 1005–1015 (2012).

92. Moura, A. *et al.* Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes. Nature Microbiology* **2**, 16185 (2016).

93. Cody, A. J. *et al.* Real-time genomic epidemiological evaluation of human *Campylobacter* isolates by use of whole-genome multilocus sequence typing. *Journal of Clinical Microbiology* **51**, 2526–2534 (2013).

94. Cody, A. J., Bray, J. E., Jolley, K. A., McCarthy, N. D. & Maiden, M. C. Core genome multilocus sequence typing scheme for stable, comparative analyses of *Campylobacter jejuni* and *C. coli* human disease isolates. *Journal of Clinical Microbiology,* JCM–00080 (2017).

95. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* **102**, 13950–13955 (2005).

96. Stahl, M. & Stintzi, A. Identification of essential genes in *C. jejuni* genome highlights hyper-variable plasticity regions. *Functional & Integrative Genomics* **11**, 241–257 (2011).

97. Lefébure, T., Pavinski Bitar, P. D., Suzuki, H. & Stanhope, M. J. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. *Genome Biology and Evolution* **2**, 646–655 (2010).

98. Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal *Escherichia coli. Nature Reviews Microbiology* **8**, 207–217 (2010).

99. Koskiniemi, S., Sun, S., Berg, O. G. & Andersson, D. I. Selection-driven gene loss in bacteria. *PLoS Genetics* **8**, 1–7 (June 2012).

100. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 1 (2009).

101. Zhao, Y. *et al.* PGAP: pan-genomes analysis pipeline. *Bioinformatics* **28**, 416–418 (2012).

102. Fouts, D. E., Brinkac, L., Beck, E., Inman, J. & Sutton, G. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Research* **40**, e172–e172 (2012).

103. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

104. Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to identify similar regions in large sequence sets. *Current Protocols in Bioinformatics,* 10–3 (2003).

105. Laing, C. *et al.* Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics* **11**, 1 (2010).

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

107. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

108. Wassenaar, T. M., Fry, B. N. & Van der Zeijst, B. A. Variation of the flagellin gene locus of *Campylobacter jejuni* by recombination and horizontal gene transfer. *Microbiology* **141**, 95–101 (1995).

109. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665 (2000).

110. Nichols, G. L., Richardson, J. F., Sheppard, S. K., Lane, C. & Sarran, C. *Campylobacter* epidemiology: a descriptive study reviewing 1 million cases in England and Wales between 1989 and 2011. *BMJ Open* **2**, e001179 (2012).

111. Sheppard, S. K. *et al. Campylobacter* genotyping to determine the source of human infection. *Clinical Infectious Diseases* **48**, 1072–1078 (2009).

112. Tam, C. C. *et al.* Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut* (2011).

113. Baker, M. G., Sneyd, E. & Wilson, N. A. Is the major increase in notified campylobacteriosis in New Zealand real? *Epidemiology & Infection* **135**, 163–170 (2007).

114. Tam, C. C. & O'Brien, S. J. Economic cost of *Campylobacter*, Norovirus and Rotavirus disease in the United Kingdom. *PLoS One* **11**, e0138526 (2016).

115. Buzby, J. C., Allos, B. M. & Roberts, T. The economic burden of *Campylobacter*-associated Guillain-Barré syndrome. *Journal of Infectious Diseases* **176**, S192–S197 (1997).

116. Scharff, R. L. Economic burden from health losses due to foodborne illness in the United States. *Journal of Food Protection* **75**, 123–131 (2012).

117. Sheppard, S. K., Jolley, K. A. & Maiden, M. C. A gene-by-gene approach to bacterial population genomics: whole genome MLST of *Campylobacter*. *Genes* **3**, 261–277 (2012).

118. Dingle, K. *et al.* Multilocus sequence typing system for *Campylobacter jejuni*. *Journal of Clinical Microbiology* **39**, 14–23 (2001).

119. SRA Toolkit Development Team. *SRA Toolkit* Dec. 2017. <https://ncbi.github.io/sra-tools>.

120. Machado, M. P. *INNUca* Dec. 2017. <https://github.com/B-UMMI/INNUca>.

121. Nurk, S. *et al. Assembling genomes and mini-metagenomes from highly chimeric reads* in *Research in Computational Molecular Biology* (2013), 158.

122. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

123. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**, R46 (2014).

124. Kruczkiewicz, P. *et al. MIST: a tool for rapid* in silico *generation of molecular data from bacterial genome sequences* in *Proceedings of the International Conference on Bioinformatics Models, Methods, and Algorithms. BIOSTEC 2013* (2013), 316–323.

125. Aho, A. V., Kernighan, B. W. & Weinberger, P. J. *The AWK programming language* (1987).

126. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. & Crook, D. W. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics* **13**, 601 (2012).

127. R Core Team. *R: a language and environment for statistical computing* R Foundation for Statistical Computing (Vienna, Austria, 2017). <https://www.R-project.org>.

128. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

129. Severiano, A., Pinto, F. R., Ramirez, M. & Carriço, J. A. Adjusted Wallace coefficient as a measure of congruence between typing methods. *Journal of Clinical Microbiology* **49**, 3997–4000 (2011).

130. Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 86–97 (2012).

131. Bayes, T. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions,* 370–418 (1763).

132. Van Rossum, G. & Drake, F. L. *The Python language reference manual* (Network Theory Ltd., 2011).

133. Van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* **13,** 22–30 (Mar. 2011).

134. McKinney, W. *Data structures for statistical computing in Python* in *Proceedings of the 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) (2010), 51–56.

135. Chao, A. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11,** 265–270 (1984).