University of Oxford



Patterns of genetic diversity in clonally evolving bacteria

Sofia Hauck
St Cross College

Department of Zoology

A thesis submitted for the degree of

Doctor of Philosophy

October 2017

Patterns of genetic diversity in clonally evolving bacteria

Sofia Hauck, St Cross College
DPhil Thesis, Trinity Term 2017

Abstract

All bacteria reproduce asexually, and most rely on the transfer of genetic material by parasexual means to adapt through the acquisition of novel genotypes and to remove deleterious mutations. A handful of bacterial species do not undergo any such horizontal gene transfer (HGT), and evolve in a fully clonal manner. They gain novel genotypes solely through the accumulation of de novo mutations, and can only remove deleterious mutations by genome-wide selective sweeps. These clonally evolving bacteria include some of the most pathogenic species known to humans.

As a consequence of their restricted means of increasing diversity and frequent purges of variation due to genome-wide sweeps, these species have very limited genetic diversity. The low resolution of the earliest molecular typing methods led to the members of these clonal species being considered identical clones, whose genetic diversity was negligible. The advent of whole genome sequencing (WGS) has revealed this assumption to be incorrect, by providing high-resolution characterisation of their genomes and revealing their existing diversity. The detailed data provided by WGS has so far been used primarily for tracking transmission and detecting drug resistance mutations, but it has potential as the basis for an exploration of population-level patterns in diversity and evolutionary selection.

In this thesis, I develop a method of rapid and scalable analysis of these patterns that is suitable to large datasets, with thousands of genomes and thousands of genes. This allows for the full diversity of the population to be considered at once. I apply this method extensively to *Mycobacterium tuberculosis*, the best-studied clonal bacterial species and a widespread human pathogen. Lastly I compare the patterns in *M. tuberculosis* to three bacterial species which do undertake HGT, *Campylobacter jejuni*, *Staphylococcus aureus* and *Neisseria meningitidis*, to elucidate the patterns specific to clonally evolving bacteria.

Thesis: ***Table of Contents***

Abstract 1

Acknowledgments 3

Note on format and authorship 4

General Introduction 5

Chapter I — Clonally evolving pathogenic bacteria 14

Chapter II — GUAVA: Rapid analysis of sequence diversity and selection in gene-by-gene annotated data sets 43

Chapter III — Lineages of Mycobacterium tuberculosis differ in their patterns of genetic diversity 75

Chapter IV — Patterns of genetic diversity and evolutionary selection in the core genomes of four bacterial species 135

General Discussion 168

Acknowledgments

I am grateful to my supervisor, Prof. Martin Maiden, for his guidance and encouragement over my career as a scientist, both in this project and in the later stages of my undergraduate degree. None of this would have been possible without his support.

I will also be forever in debt to the remainder of my the Maiden Lab group. The warm camaraderie of the DPhil students will stay with me for the rest of my life, and for that I must thank Holly, Helen, Ellie, Dot, Melissa, Kanny, Clare, Maddi, Marianne, Adedoyin and Charlene. I will always cherish the discussions, on academic subjects and all others besides, I have had with the researchers in this group, and so I express my gratitude to Odile, Noel, Jenny, Fran, Ali, Mingliang, Keith, Carina, Kenneth, and especially to James.

I must also thank the institutions that financially supported me during this project. First of all, the Conselho Nacional de Desenvolvimento Científico e Tecnológico, or CNPq, for the stipend that supported my living costs. My college and department, namely St Cross College and the Department of Zoology, and the EU Consortium Patho-NGen-Trace project provided travel funds that enabled me to travel all over the world and attend conferences pertinent to this project. The chance to meet and hear from the myriad of incredible people working in the fascinating nexus of public health, bioinformatics and evolutionary biology was invaluable for my growth as a scientist.

I owe my gratitude to my family: my parents Miguel and Erica, and my sister Helena, for their support. I am similarly in debt to Chris, and the countless friends I have made at St Hugh's, St Cross, Wolfson and every other college in the unforgettable city of dreaming spires.

Finally, I thank my examiners, for their patience and invaluable feedback.

Note on format and authorship

This thesis is comprised of six sections. The first section, in which this page is found, also contains the Abstract, a Table of Contents, the Acknowledgments and the General Introduction. The next four sections make up the bulk of the thesis, in the form of four chapters. The first chapter, titled "Clonally evolving pathogenic bacteria" is a review that was published in 2018 as an article in the book *Molecular Mechanisms of Microbial Evolution*. The remaining three chapters have not yet been published, but are presented here in article manuscript format. The second chapter, titled "GUAVA: Rapid analysis of sequence diversity and selection in gene-by-gene annotated data sets," describes the software written by me in order to perform the analyses undertaken in the following two chapters. Chapter III, titled "Lineages of Mycobacterium tuberculosis differ in their patterns of genetic diversity", is an in-depth investigation of genetic signals in a large collection of *M. tuberculosis* genomes. The last chapter, titled "Patterns of genetic diversity and evolutionary selection in the core genomes of four bacterial species", is a wider application of the same method. The sixth section contains a general discussion and general concluding remarks. Each section is presented with its own Table of Contents, near its beginning, and own list of References, near its end. Page numbers are given in the footer, in both the page number cumulatively in the thesis document and the page number out of the total page counter for the section.

All four chapters consist of work that has been conceived, designed and performed by me, including all analysis and interpretation of data. I created and maintained the data sets used in Chapters II and III. Chapter IV lists details of the provenance of its data. I wrote all parts of these manuscripts. Other authors contributed software and guidance on its use (KAJ), assembled genomes and coding guidance (JEB), and comments on manuscripts (JEB and MCJM).

Corrections submitted Michaelmas Term 2018.

General Introduction

Bacteria are single-celled organisms, but their small size does not hamper their ability to kill humans by the millions. Bacteria contribute to three of the top ten causes of death in 2015, with tuberculosis alone coming into 9th place, causing the equivalent of three deaths a minute. This high disease burden represents a marked improvement over the even higher rates of bacterial disease observed in the last century, following from millennia of human suffering caused by bacteria (World Health Organization 2015).

Tuberculosis disease had "ravaged the heart of Europe" for thousands of years before Robert Koch first formally described *Mycobacterium tuberculosis*, its etiological agent, in Berlin in the year 1882. During Koch's lifetime, it was responsible for one in seven deaths in central Europe (Kay et al. 2015). In the decades that followed, scientific research turned the tide against the tyranny of tuberculosis: the BCG vaccine was first used in 1921, and antibiotics that could treat *M. tuberculosis* infections were first applied during the second world war, merely decades after Koch's reports. The rates of tuberculosis cases and deaths plummeted following these advances. Ridding humanity of tuberculosis altogether seemed an achievable goal (Luca and Mihaescu 2013).

Today this dream has been crushed by the emergence and astounding rise in prevalence of drug resistant tuberculosis. Strains of *M. tuberculosis* resistant to the both first-choice drugs used against it, rifampicin and isoniazid, are now found worldwide. They are known as multi-drug resistant tuberculosis, or MDR-TB, and are responsible for around half a million new cases each year. MDR-TB cases cost 25 times more to treat than those due to drug-susceptible TB, and the treatment succeeds only two-thirds as often as it does for susceptible infections. These factors, combined with the continuing difficulty in the timely identification of drug resistance, has led to a growing epidemic of MDR-TB (World Health Organization 2016). The threats do not end there, as

the need for the increasingly frightening acronyms points to: there is now XDR-TB, extensively drug resistant tuberculosis, invulnerable to both first- and second-choice drugs, and TDR-TB, totally drug resistant tuberculosis, which is resistant to a shocking thirteen classes of antibiotic drugs (Jang et al. 2017). The high rates of co-infection with HIV, especially in developing countries, conflates the tuberculosis and AIDS epidemics, complicating the control of both diseases (Marais 2016).

It is worth noting that these developments are not unique to tuberculosis. Cholera was first recorded as a disease in antiquity, simultaneously in ancient Greece and ancient India. *Vibrio cholerae*, the bacteria responsible for cholera, was the cause of major epidemics worldwide throughout the 18th, 19th and 20th centuries. Famously, during a cholera outbreak in London in 1849, Dr. John Snow performed what might be one of the first epidemiological studies. He mapped the location of cholera deaths, which combined with his knowledge of the water-borne transmission of the bacteria, pointed to a contaminated water pump as the source of the outbreak (Colwell 1996). While the epidemic waves of cholera today are smaller than they have been in the past, cholera remains a large public health threat around the world, and one that is likely to grow, as the severity of its epidemics increases with globalisation and climate change (Baracchini et al. 2017).

Syphilis is yet another disease that follows this historical pattern. It has been known since the 14th century within Europe, and may have circulated for far longer in the New World. The advent of the global shipping trade spread it around the globe and it became a major human pathogen. Technological advances in the early 20th century in both microscope-based diagnostic tests and drugs such as penicillin finally brought *Treponema pallidum*, the bacteria that causes syphilis, under control (Maatouk and Moutran 2014). The last few decades have seen a reversal in this progress, as dramatic increases in the prevalence of syphilis can now be observed in even in

high-income countries. The rising rates of antibiotic resistant syphilis, HIV co-infection and global shortages of the drugs used for treatment have caused outbreaks of *Treponema* to take hold once again (Hook 2017).

A pattern emerges from these examples: bacterial diseases have in many cases been known for centuries, and were especially prevalent in the last century or two. They were all brought to heel in the 20th century by advances in diagnostics, prevention and treatment, so that their incidence and mortality were dramatically lowered. However, the promise of these new technologies has not been borne out to the point of eradication. Instead, bacterial diseases are responsible for resurging disease burdens in the 21st century. Complications such as growing drug resistance, co-infection with other pathogens, and climate change have raised questions about whether their management is possible in the future.

The surprising adaptability of these single-celled organisms lends them the ability to circumvent the best controls the human medical and scientific establishment has to offer. The balance in the host-pathogen arms race is shifting back towards bacteria. To reverse this trend and keep the threat posed by these microbial pathogens at bay, we must consider the context in which these pathogenic bacteria arise.

Bacteria represent one of the three domains of life, alongside Archaea and Eukaryota. It is possibly the oldest domain, with bacterial fossils dated as 3,770 million years old pointing to their emergence as likely the very first life forms on earth (Dodd et al. 2017). Since then they have expanded, reproducing by duplication and diverging by natural selection so as to adapt to every imaginable niche available on Earth. Bacterial species vary tremendously in every aspect of their lifestyle, metabolism and genetics. Metagenomic studies that can capture the genetic

sequences of bacteria without the need to culture them in a laboratory beforehand have found that an immerse diversity of bacteria in a variety of environments (Hug et al. 2016). Our own bodies are host to a myriad of bacterial species living in our "microbiome" (Bashan et al. 2016). Only a tiny minority of this diverse domain is pathogenic. Pathogens have, however, emerged from all across the bacterial domain independently, and so can provide an overview of the domain as a whole (Maiden et al. 2013). The evolutionary drivers of pathogenicity in bacteria, and their independent occurrence throughout the bacterial domain, are explored further in Chapter I.

One of the fundamental questions in the study of the bacterial population genetics is centered on the ability of bacteria to horizontally, or laterally, transfer genetic material to one another. Parasexual means of genetic recombination lend bacteria the ability to evolve quickly, by giving them the ability to acquire beneficial genotypes, such as those that grant resistant to antibiotics, from almost all parts of the domain (Cohan 2002). At what rate this horizontal gene transfer occurs, and what factors cause this rate to vary, are questions we are only beginning to answer. We do know that this kind of genetic recombination cannot be so frequent so as to break down all barriers in the bacterial domain, as the same clusters of genes are discovered repeatedly, throughout the world and whenever bacteria are sampled (Hanage et al. 2006). In other words, genetic diversity in bacteria is "not continuous", but instead, "distinct, stable types with particular properties exist" (Maiden et al. 2013), as observed in both phenotypes and genotypes. How the balance between horizontal gene transfer and vertical, or clonal, gene transfer is determined and how its variation affects bacterial populations has remained a major area of research for decades (Maynard Smith et al. 1993; Shapiro 2016).

The curious existence of fully clonal bacterial species therefore presents a crucial model from which the evolution of bacteria can be investigated (Achtman 2012). Fully clonal bacteria do not undergo any HGT, existing as the extreme of a spectrum that contains a variety of evolutionary patterns. Understanding of the effect of the lack of HGT can therefore serve as a critical foothold for the study of the bewildering diversity within the bacterial domain.

As covered in more detail in Chapter I, these clonal bacteria rely solely on mutations that occur randomly in their own genomes to increase the genetic diversity of their populations. It may be easy to presume from this limitation that clonal bacteria would therefore be woefully underprepared for the arms race between germ and host that defines co-adaptation in a pathogenic relationship. This is a false assumption; the clonally evolving bacteria include many of the most pathogenic species known, including the previously discussed *Mycobacterium tuberculosis*.

Along with the limited methods of gaining new variation, populations under HGT are limited in their means of removing detrimental variation. When no recombination is possible, the entire genome must evolve as one unit, by genome-wide sweeps that purge the diversity in all genes even if only a single gene is being strongly selected for (Smith et al. 2006). Because of this low diversity, clonal populations were sometimes considered to consist of identical clones, whose diversity that was either non-existent or meaningless (Gagneux and Small 2007).

Next-generation sequencing technologies encompass vast improvements to the speed, cost and accuracy of whole genome sequencing. The widespread use of whole genome sequencing that followed from this has completely changed the scale at which investigations into the population genetics of bacteria can be carried out. The increased resolution afforded by whole genome sequencing has shattered the preconception of clonal populations as containing only identical clones. The use of the term “clonal” to refer to these populations therefore describes

their clonal (as opposed to sexual) reproduction, not level of their genetic diversity, as explained further in Chapter I.

The number of whole genome sequences available for bacteria is now in the hundreds of thousands, but analysis tools have not kept pace (Coscolla and Gagneux 2014). Indexing the diversity within a population by the alleles present, in what can be called an allele-based or gene-by-gene approach, is one way to circumvent this obstacle to large-scale analyses (Maiden et al. 1998). They are based on the identification of unique alleles, versions of genetic sequences, which are recorded by the use of an identifier in the form of an arbitrary integer. While summarising genetic sequences into a single identifier does entail a loss of resolution and a resultant loss in the power of analyses, the uniqueness of identifiers means that the full resolution can be recovered if necessary (Didelot and Falush 2007). Because allele-based methods generally scale in a linear rather than exponential fashion, they can remain relevant in a context where data itself is increasing exponentially (Maiden and Jolley 2010). Considering the whole gene at once, instead of individual sites, also allows for a simple correction of the effect of recombination. HGT can cause an entirely novel allele to be transferred in a single genetic event, not to mention a wide variety of sub-allelic variation. Weighing an HGT event as equivalent to one de novo mutation brings diversity caused by mutation and caused by gene transfer in line with one another in terms of measuring evolutionary "events", as opposed to the impact on the genome. It also eliminates the need to categorise mutations as occurring due to mutation or due to recombination as is often required for alignment-based methods of phylogenetic analysis (Maiden 2006).

These features of the gene-by-gene model allow it to be applied to very large data sets, with tens of thousands of isolates and thousands of loci. This thesis presents the development of

software that uses the model to analyse such very large data sets (in Chapter II), and its application for both in-depth analysis of one population (in Chapter III) and comparatively for multiple species (in Chapter IV).

References cited in General Introduction

Achtman M (2012) Insights from genomic comparisons of genetically monomorphic bacterial pathogens. Phil Trans R Soc B 367:860–867. doi: 10.1098/rstb.2011.0303

Baracchini T, King AA, Bouma MJ, et al (2017) Seasonality in cholera dynamics: A rainfall-driven model explains the wide range of patterns in endemic areas. 1–10. doi: 10.1016/j.advwatres.2016.11.012

Bashan A, Gibson TE, Friedman J, et al (2016) Universality of human microbial dynamics. Nature 534:259–262. doi: 10.1038/nature18301

Cohan FM (2002) Sexual isolation and speciation in bacteria. Genetica 116:359–370. doi: 10.1023/A:1021232409545

Colwell RR (1996) Global climate and infectious disease: The cholera paradigm. Science 274:2025–2031.

Coscolla M, Gagneux S (2014) Consequences of genomic diversity in Mycobacterium tuberculosis. Semin Immunol 26:431–444. doi: 10.1016/j.smim.2014.09.012

Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–1266. doi: 10.1534/genetics.106.063305

Dodd MS, Papineau D, Grenne T, et al (2017) Evidence for early life in Earth's oldest hydrothermal vent precipitates. Nature 543:60–64. doi: 10.1038/nature21377

Gagneux S, Small PM (2007) Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. Lancet Infect Dis 7:328–337. doi: 10.1016/S1473-3099(07)70108-1

Hanage WP, Fraser C, Spratt BG (2006) Sequences, sequence clusters and bacterial species. Phil Trans R Soc B 361:1917–1927. doi: 10.1098/rstb.2006.1917

Hook EW 3rd (2017) Syphilis. Lancet 389:1550–1557. doi: 10.1016/S0140-6736(16)32411-4

Hug LA, Baker BJ, Anantharaman K, et al (2016) A new view of the tree of life. Nature Microbiology. doi: 10.1038/NMICROBIOL.2016.48

Jang J-C, Jung Y-G, Choi J, et al (2017) Bedaquiline susceptibility test for totally drug-resistant tuberculosis Mycobacterium tuberculosis. J Microbiol 55:483–487. doi: 10.1007/s12275-017-6630-1

Kay GL, Sergeant MJ, Zhou Z, et al (2015) Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nature Communications 6:6717. doi: 10.1038/ncomms7717

Luca S, Mihaescu T (2013) History of BCG Vaccine. Maedica (Buchar) 8:53–58.

Maatouk I, Moutran R (2014) History of syphilis: between poetry and medicine. J Sex Med 11:307–310. doi: 10.1111/jsm.12354

Maiden M, Bygraves JA, Feil E, et al (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145. doi: 10.1073/pnas.95.6.3140

Maiden MCJ (2006) Multilocus sequence typing of bacteria. Microbiol Immunol 60:561–588. doi: 10.1146/annurev.micro.59.030804.121325

Maiden MCJ, Jolley KA (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595. doi: 10.1186/1471-2105-11-595

Maiden MCJ, van Rensburg MJJ, Bray JE, et al (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:728–736. doi: 10.1038/nrmicro3093

Marais BJ (2016) The global tuberculosis situation and the inexorable rise of drug-resistant disease. Adv Drug Deliv Rev 102:3–9. doi: 10.1016/j.addr.2016.01.021

Maynard Smith J, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? Proc Natl Acad Sci USA 90:4384–4388. doi: 10.1902/jop.2011.110063

Shapiro BJ (2016) How clonal are bacteria over time? Current Opinion in Microbiology 31:116–123. doi: 10.1016/j.mib.2016.03.013

Smith NH, Gordon SV, la Rua-Domenech de R, et al (2006) Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat Rev Microbiol 4:670–681. doi: 10.1038/nrmicro1472

World Health Organization (2016) Global Tuberculosis Report 2016.

World Health Organization (2015), Top 10 Causes of Death in 2000 and 2015. http://www.who.int/mediacentre/factsheets/fs310/en/.

Chapter I
Clonally Evolving Pathogenic Bacteria

Sofia Hauck and Martin C J Maiden

Department of Zoology, University of Oxford, Oxford, United Kingdom

This chapter is reproduced here as published in the book Molecular Mechanisms of Microbial Evolution (Grand Challenges in Biology and Biotechnology), edited by Pabulo H. Rampelotto, Springer 2018, doi: 10.1007/978-3-319-69078-0

Abstract

Unlike those of animals and plants, sexual processes in bacteria are not linked to reproduction *per se* and are a consequence of Horizontal (sometimes called Lateral) Genetic Transfer (HGT). This is the acquisition of genetic material from a non-parental source and its incorporation into the genome, usually by recombination. The relative and absolute rates of HGT in different bacteria vary greatly, and this has a major influence on population diversity, structure, and evolution. Rates of HGT are influenced by both mechanistic and ecological considerations and when HGT is absent or very rare, bacterial populations become completely clonal. In such populations, evolution occurs solely by the accumulation of mutations, accompanied by diversity reduction events such as periodic selection and/or stochastic bottlenecks, which can result in hitchhiking events that fix mutations, whether beneficial, deleterious, or neutral in their effect on fitness. The net effect is the establishment of distinct lineages within the population, which are characterized by low genetic diversity accompanied by the relaxation of purifying selection and an enhancement of the effects of

genetic drift. The best examples of such bacteria are well-known obligate pathogens that affect humans, including *Mycobacterium tuberculosis, Mycobacterium leprae, Yersinia pestis*, *Bacillus anthracis*, *Salmonella enterica* serovar Typhi, and *Bordetella pertussis.* Each of these pathogens has evolved from a more diverse parental population as a result of a shift in niche from a free-living organism to an obligate parasite. In some cases, this was caused or accompanied by a HGT event, but ultimately the process has led to reproductive isolation: the inhibition of further HGT with the parental population, the wider bacteria community, and even among members of the population itself. While this reproductive isolation may have promoted specialization and eventually speciation, it can also result in loss of fitness, as a consequence of the accumulation of deleterious or mildly deleterious mutations, a process known as Muller's Ratchet. Since their emergence, the evolution of these organisms has been further influenced by ecological factors resulting from changes in human behavior, including increased population numbers and density, and the use of antibiotics.

Chapter I: ***Table of Contents***

Abstract **1**

1. Clonality in Bacteria **3**

2. The Clonal Paradigm **4**

3. Limiting the Rate of Recombination **10**

4. Mutation and the Evolution of Clonal Pathogens **13**

5. Examples of Adaptive Evolution in Clonal Pathogens **16**

6. Consequences of Clonality **18**

Figure 1: Illustration of bacterial reproduction and evolution, including epidemic clonality and fully clonal evolution. 22

References **24**

1. Clonality in Bacteria

A population is a group of organisms that occupy the same niche and that share an evolutionary path, adapting to their environment through the process of natural selection. Members of a population are closely related and united by a recent common ancestor or ancestral population. Despite these similarities, bacterial populations are rarely composed of single clones that are genetically identical. Indeed, genetic variation within a population is a necessary condition for evolution by means of natural selection to occur. A population of identical clones cannot adapt, as all of its members have genomes that are equally suited to any set of conditions (Nielsen 2005).

Natural selection results in changes to the frequency of genotypes according to their fitness, as higher rates of reproduction and survival increase the abundance of the fittest genotypes within the population. Likewise, lower rates of success reduce the frequency of the least fit genotypes, sometimes to the point of elimination. This process therefore reduces the genetic diversity of the population over time, as it can only maintain or decrease diversity. Diversity is not completely reduced by natural selection because it is balanced by the introduction of new variants through mutation and recombination. All genetic variation ultimately arises from mutations, but sexual processes can introduce mutations from other populations via horizontal gene transfer (HGT) as well as causing the re-assortment of genes within a population (Feil and Spratt 2001).

Sexual recombination in bacteria differs from that of plants and animals in both frequency and range of source genetic material. Bacterial HGT occurs widely and promiscuously, with genetic material sourced throughout the domain, rather than exclusively from other members of the same species (Cohan 2002). Unlike most plants and animals, which recombine as a matter of course during each reproductive cycle, HGT in bacteria occurs at rates and extents that vary widely by species and by ecological situation (Hanage et al. 2006a). Across the bacterial domain, a wide spectrum of recombination rates has been

observed, from very high rates that erase most clonal signal to complete clonality (Spratt and Maiden 1999). In a fundamentally non-clonal population, genes are the evolutionary units, as recombination is so frequent that it overwhelms the non-random associations among genetic variants in a genome, collectively named "linkage disequilibrium". A recombination rate that is ten times higher than the mutation rate may be enough for the linkage caused by shared descent to break down entirely (Cohan 1994). When this is the case, each variant of a particular gene can exist in a range of genomic backgrounds, and therefore evolve independently. This is not a hypothetical extreme as it occurs in *Helicobacter pylori*, a human-specific pathogen that causes chronic gastric infections (Suerbaum et al. 1998), which has been found to exchange up to a tenth of its genome in one multi-strain infection over a period of four years (Cao et al. 2015).

At the opposite end of the spectrum is full clonality, which is a consequence of limited or no HGT (Tibayrenc and Ayala 2002). HGT can occur occasionally without affecting a clonal population structure, as long as it is infrequent enough to remain beneath the "clonality threshold". This is the point at which recombination is too rare to act as a homogenising force within the population and prevent the divergence of distinct branches (Tibayrenc and Ayala 2015). In the extreme and unusual case of a fully clonal population, where no recombination can be observed, genes cannot be exchanged among individuals and therefore each genome evolves independently. The phylogeny of a population in these conditions fits the "Russian Doll" model, where with increasing resolution, each branch of the species phylogeny is subdivided into smaller distinct and diverging branches (Tibayrenc and Ayala 2016).

2. The Clonal Paradigm

This article will focus on fully clonal populations, where HGT is entirely absent, or sufficiently rare that it has essentially no impact on the evolution, population structure, and ecology of the organism in question during the timeframe under consideration (Shapiro

2016). The first major consequence of limited or no HGT is the inability to repair mutations that arise from time to time. Bacteria can use a variety of mechanisms for repairing DNA damage, including nucleotide or base excision repair and mismatch repair. These mechanisms can lower the rate of mutation by reversing damage as it happens, but without HGT, bacteria cannot repair mutations that have been copied to both stands of DNA (van der Veen et al. 2015). Two other major consequences follow from a lack of HGT: the absence of a mechanism whereby different mutations can be combined to make novel variants and the inability to acquire genetic material from other organisms. All of three of these affect the diversity of a population. In the absence of HGT, a population is both limited in its ability to generate new genetic diversity and to maintain existing diversity, as it is frequently subject to purges (Namouchi et al. 2012) by means of bottlenecks when population sizes are small or genome-wide sweeps when selection is strong (Smith et al. 2006). This lack of diversity has consequences for the adaptability and long-term viability of a population.

Populations that do not participate in HGT typically exhibit very low genetic diversity, although a lack of diversity is also a feature of relatively young populations, as over time mutations will accumulate even in the absence of HGT. The term "clonal" can describe both of these traits, the genealogical concept of reproduction through clonal descent and the genetic concept of identical clones forming all members of a population. This state of low diversity within the population is referred to as genetic monomorphism. The confusion between these definitions is particularly pronounced because these two traits often, but not necessarily, co-exist in bacteria (Tibayrenc and Ayala 2012).

From the perspective of evolutionary and molecular epidemiological studies, the low diversity of monomorphic bacteria can present problems for the identification of epidemiologically relevant variants for practical purposes. When typing resolution is not sufficient to distinguish similar isolates, the crucial task of tracking pathogen transmission becomes more challenging (Kohl et al. 2014). Improvements in the speed, cost and accuracy

of whole genome sequencing (WGS) made in the 10 years prior to writing this article, ensured that these bacteria could be examined at an unprecedented level of detail, covering virtually all of the existing genetic variation. Consequently, fully clonal, monomorphic pathogen species have been some of the first to be analysed in-depth by this technology, precisely because their low diversity simplifies large-scale genomics studies (Achtman 2008). This research has made it possible to discover the common traits of fully clonal bacteria.

Most known examples of highly clonal pathogens are lineages nested within a more diverse phylogeny of non pathogenic organisms (Achtman 2008). An example of this is the agent of tuberculosis, *Mycobacterium tuberculosis*, which is perhaps the best studied example of a fully clonal bacterium. It is one of nine lineages within the *Mycobacterium tuberculosis* complex, each of which is host-specific, though some spillover in hosts does exist (Ghodbane et al. 2014). The complex as a whole is considered to be one species, due to its high degree of relatedness (99% or more average nucleotide identity), but individual species names are still used to refer to each subspecies, or ecotype, within it (Djelouadji et al. 2011). There are a number of parallel examples elsewhere in the domain including: *Yersinia pestis*, the agent of plague, which is best considered a specific clone of *Yersinia pseudotuberculosis* as it is nested within this more diverse species (Rasmussen et al. 2015), *Bordetella pertussis,* which causes whooping cough, is a monophyletic branch of the *B. bronchiseptica* species (Diavatopoulos et al. 2005), and *Bacillus anthracis,* responsible for anthrax, is an equivalent branch of the *B. cereus* phylogeny (Okinaka et al. 2006). Not all clonal populations are given species names, as can be observed with *Salmonella enterica* subspecies *enterica* serovar Typhi, the agent of typhoid fever (Didelot et al. 2011).

In the absence of wholly sexual reproduction, which is used to define animal species, and the widespread occurrence of HGT in bacteria from a variety of sources, precise definitions of bacterial species remain difficult, contingent, and to a degree controversial (Achtman and Wagner 2008). For pathogenic bacteria, the importance of accurately circumscribing

disease-causing microbes has led to pathogenicity being used as the defining trait of a species. However, pathogenicity, like many other phenotypic traits, correlates poorly with genotype, as is becoming increasing clear as genotyping methods improve. A species name that may not be "accurate", that is, not defining a distinct monophyletic cluster of genotypic and phenotypic traits, may nontheless still be used in cases where distinguishing dangerous strains from harmless genetic relatives can be a life-or-death concern. This concept of a *nomen periculosum* is one example of the many complications in the field of bacterial taxonomy (Stackebrandt et al. 2002).

The complex nature of phenotypic and genotypic clusters in bacteria, both in defining clear clusters by either set of criteria or in correlating genotype to phenotype, has meant that a universal bacterial species concept is yet to be achieved. While the search for a unifying definition continues, "species" remains a term without one particular meaning, largely used as a label of convenience (Bapteste and Boucher 2009). Lineages are given names "dependent on the level of divergence thought appropriate" and only "where this serves a useful purpose" (Lan and Reeves, 2001). Clonal populations present particular problems in defining species clusters, as the absence of recombination means that each lineage is "irreversibly set" on divergence from other members of its species (Hanage et al. 2006b). For this reason, "species", "clone", "serovar" or any other terms for a specific phylogenetic group of bacteria are used here as per the convention for that species.

Despite HGT not being a mechanism in the maintenance of clonal lineages, it is often key to the evolutionary events that have led to them becoming distinct from their parental populations. HGT events can be catalysts for the expansion of a population to a new niche, and form critical junctions in the evolutionary path of a species (Wiedenbeck and Cohan 2011). A clear example of this has been observed in the history of *Bacillus anthracis,* the agent of anthrax. It diverged from *Bacillus cereus* after acquiring the pXO1 and pXO2 plasmids, which includes the anthrax toxin genes implicated in the high fatality rate of the disease

(Kolstø et al. 2009). After this acquisition, which occurred perhaps 20,000 years ago, *B. anthracis*'s unique pathobiology has caused it to specialize into a novel ecological niche (Keim et al. 2009). The evolutionary history of *S. enterica* serovar Typhi displays a similar pattern. Typhi lost the promiscuous HGT that is characteristic of the *S. enterica* species, changed niche as it transformed from a gastroenteric infection to an invasive one and underwent functional gene loss and genome degradation in the process (Holt et al. 2008).

Irreversible specialisation is the only possibility for fully clonal populations, due to the loss of adaptability that comes with restricted HGT (Moran and Wernegreen 2000). Parasitism itself is a form of specialisation, and the extreme host-specificity and obligate pathogenicity evident in many fully clonal species is an even more extreme version of this specialisation (Shapiro 2016). While the possibility of a discovery bias caused by the overrepresentation of pathogens in bacterial research cannot be discounted, it appears that all persistent fully clonal bacteria are pathogenic (Achtman 2012).

Many of the most deadly bacterial pathogens are species with low genetic diversity that fit the fully clonal paradigm (Namouchi et al. 2012). For example, the etiological agent of plague, *Y. pestis*, was responsible for the Black Death pandemic that killed nearly a third of Europeans in the 14th century, and this organism remains the cause of small but often fatal outbreaks in the modern era (Gage and Kosoy 2005). Meanwhile, *M. tuberculosis* is the most deadly bacterial pathogen in the world, infecting 100 million new patients and causing 1.5 million deaths a year, and appears to have coevolved with humans (Niemann et al. 2016). *S. enterica* serovar Typhi causes 200,000 deaths a year and may have been the cause of the plague of Athens in 430 B.C. (Galan 2016).

These species are genetically monomorphic despite their wide geographic range and long history as human pathogens, stretching over tens of thousands of years. While this is not substantial stretch of time in evolutionary terms, it is enough for analyses of mutation rate to be made across species, and for phylogeographic signals to emerge within species (Comas et

al. 2013). A comparison between modern *Y. pestis* strains and DNA recovered from a Black Death victim showed that few genetic changes occurred in the interceding 660 years, with only 10 cases of altered gene order and only 97 chromosomal sites and 6 plasmid single-nucleotide sites differing from modern strains (Bos et al. 2011). For *M. tuberculosis*, whole genome alignments find that even the most distant strains that can be isolated from humans differ at most in 1,800 single nucleotide polymorphisms, or SNPs (Coscolla and Gagneux 2014). In contrast, over 15,000 SNP differences have been found between isolates of *Staphylococcus aureus* (Planet et al. 2017) and *Campylobacter jejuni* isolates can differ by 13,000 SNPs even within one outbreak (Moffatt et al. 2016).

Lifestyle stages with slow or no growth, common to many clonal pathogenic bacteria, may further exacerbate their low genetic diversity, by prolonging generation times and slowing evolutionary rates (Ohta 2011). *M. tuberculosis* can enter a granuloma phase in infected individuals, resulting in essentially asymptomatic carriage (Ford et al. 2011). *Bacillus anthracis* can survive as a spore in dry soil for decades, potentially centuries, in which time it does not replicate and therefore does not accumulate mutations (Rasko et al. 2011). This does not seem to be a necessary condition, however, as *Bordetella pertussis* is an obligate human pathogen with no chronic carrier state (Park et al. 2012). The source-sink evolutionary dynamics caused by carrier stages in their lifestyles may also reduce variability by stabilizing the genome, as adaptations that are advantageous in the infectious phase confer no fitness benefits for invasive disease. This may be the case in *Yersinia pestis*, which constantly circulates in rodent populations between its infrequent outbreaks in humans (Ayyadurai et al. 2008) and in *S.* Typhi, whose reservoir is human asymptomatic carriers that may shed the bacteria for decades over their lifetimes (Holt et al. 2008).

3. Limiting the Rate of Recombination

To date, pathogens with restricted host ranges are among the best studied clonally evolving populations. A host represents a highly controlled and relatively uniform environment, and the restriction to a given host range deepens the extreme level of specialization these populations have undergone. As specialization increases, newly introduced genes are more likely to disturb existing favourable interactions in the genome, an effect named recombinational load (Michod et al. 2008). Because of this, it is advantageous for clonally evolving bacteria to actively supress their rate of HGT so as to reduce this load. The realised rate is at a maximum equivalent to the potential rate, which is determined by the available mechanisms for HGT, but which in practice is restricted by ecological conditions of the recipient bacteria (Yahara et al. 2016).

DNA can be transferred between bacteria by three mechanisms, all of which transfer a small amount of DNA from a donor to a recipient: transduction, conjugation and transformation. Briefly, transduction involves infectious virions (most often bacteriophages) as vectors, while conjugation the vectors are self-replicating plasmids or other mobile genetic elements. Bacteria can protect themselves against these parasitic elements by means of endonucleases, enzymes that cleave foreign DNA. With these enzymes, invading DNA sequences can be destroyed before or after they recombine with the bacterial chromosome and consequently prevent cases of successful HGT from being passed on to subsequent generations (Cohan 2002). Restriction endonucleases recognise and cleave DNA at specific "target" sequences, while the more recently discovered CRISPR-Cas system detects the Clustered Regulatory Short Palindromic Repeats (CRISPR) after which it is named. The high specificity of the targets in both restriction and CRISPR systems is key to their ability to protect bacteria from "foreign" DNA without damaging their own genomes.

In contrast, transformation depends on a recipient with functional and active competency machinery to complete the uptake of DNA from the environment. This requires a set of highly

specialised loci that are often spread throughout the genome and which are conserved across a species (Croucher et al. 2016). Because expressing this machinery carries a high cost, competence is typically regulated so that it is transiently expressed only under specific circumstances. When the loss of one of the genes occurs, relaxed selection on the others ensures that they are quickly lost. These losses often signal a point of no return for the lineages in which they occur, resulting in their extinction, albeit in the long-term perspective of hundreds of millions of years (Redfield et al. 2006).

Even in the presence of the possibility of HGT, for any of these mechanisms to succeed at facilitating transfer between two bacteria, the donor and recipient must be in close proximity to one another, so that the DNA may be passed from one to the other. Spatial isolation may therefore be a barrier to gene flow (Polz et al. 2013), and is the basis for the "starving sex hypothesis" that posits clonality as a passive process, occurring due to the absence of opportunity for mating (Tibayrenc and Ayala 2016). This may in fact be the case in many pathogenic bacteria, for which transmission events often mean only a small founding population infects each new host (Bergstrom et al. 1999).

The limitation of small founding populations can prevent the impact of recombination in yet another way: if the genetic distance between host and recipient is too small, the exchange of DNA is inconsequential and indeed unobservable. After DNA is exchanged, homologous recombination is a necessary step for genetic material to be incorporated into the genome and subsequently heritable. This requires that the donor DNA be paired with a chromosomal sequence with which it has some sequence identity, at least on its flanking sequences. If the donor sequence and host sequence are identical, homologous recombination may indeed be occurring, but the like-for-like exchange leaves no trace in a case of "invisible sex" (Tibayrenc and Ayala 2015). If both donor and recipient bacteria are descendants of a recent transmission event, they may be too closely related to recombine in ways that can be

observed. This may in fact be the most common form of HGT within bacteria, a mechanism that is advantageous due to its utility in DNA repair (Feil and Spratt 2001).

With the high resolution possible by WGS analyses, data from multiple samples of a single patient have shown that multi-clonal infections are more common than previously thought (Votintseva et al. 2014). For *M. tuberculosis*, DNA recovered from human remains indicates that multiple strain infections were commonplace in 18th century Europe (Kay et al. 2015). It is also likely that any colonising bacteria are not alone in their niche, as every area of the human body has an associated microbiome (Huttenhower et al. 2012); however, physical proximity to distantly related bacteria may not lead to HGT as the frequency of transformation is reduced exponentially as sequence divergence increases. Therefore, the rate of successful HGT is reduced as the genetic distance between a host and a recipient increases and the homology between their genome sequences weakens (Cohan 1994). Genetic distance between bacteria can also lower the efficacy of plasmids and bacteriophages as vectors of HGT as their host ranges are limited to similar bacteria (Cohan 2002).

In this way, recombination may not occur, even when potential mechanisms are present and the donor and recipient bacteria are physically close, if the bacteria are either too distantly related (no HGT will occur) or too closely related (HGT is non-observable). HGT is a major force in the evolution of the bacteria domain as a whole, with estimates that up to 20% of genes across the domain have been recently mobilised by this method. Bacteria rapidly adapt to changes in their environment using this method for gene flow to swap the genes in their accessory genome. In a fully clonal population, HGT is not available as a mechanism for increasing genetic diversity within a species, and so mutation remains as the only option (Feil and Spratt 2001).

4. Mutation and the Evolution of Clonal Pathogens

Mutations arise from DNA damage or copying errors and are therefore random with respect to gene function. Some gene features, such as homopolymeric tracks, may locally raise the rate of mutation, and therefore be found primarily in genes where this effect is beneficial, but even mutations such as this occur randomly (Orsi et al. 2010). According to Ohta's Nearly Neutral Theory model, the majority of mutations have little or no effect on phenotype, with a small selective coefficient if any, and out of these, most are slightly deleterious. A minority of mutations will be fatal and an even smaller number will confer a strong positive effect on fitness (Ohta 1992). Obligate pathogens, by the nature of the specialization inherent in their ecology, can be particularly likely to incur fitness costs from most mutations (Achtman and Wagner 2008). The consequence of this is that point mutations, by and large, generate mutations that negatively impact the fitness of the bacterium in which they arise.

Hyper-mutating bacteria, those that are defective in DNA proof-reading mechanisms, can develop new traits, such as antibiotic resistance, at faster rates, but this comes at the cost of their overall fitness of the organism as a whole (Woodford and Ellington 2007). This is because deleterious mutations cannot be easily removed from the genome in the absence of HGT. In strictly clonal organisms, with no HGT, linkage disequilibrium is so strong that the entire genome is the evolutionary unit and natural selection can only remove a deleterious mutation through a strong purifying selective pressure that causes a genome-wide sweep. By this mechanism, however, other deleterious mutations can become fixed in a population by hitchhiking with beneficial mutations during selective sweeps, as the sudden decrease in population size causes their mild negative effect to be counteracted by the overwhelmingly positive effect acting upon the allele selected for by the sweep (Fay and Wu 2000). Population bottlenecks, like sweeps, can remove or fix mutations; however, in this case, due to the small effective population size, it is random genetic drift, rather than natural selection, which is the

predominant force. Consequently, such bottlenecks often fix deleterious mutations that actually reduce fitness.

Distinguishing between sweeps and bottlenecks, especially in retrospect, can be difficult, due to their similar effects on population structure, in that they both cause a sudden and severe reduction in the effective population size (Smith et al. 2006). The survival of a particular variant through a bottleneck may not be entirely decided by random genetic drift when more virulent genotypes lead to higher bacterial load within the host and a consequently higher likelihood of transmission to a new host. This may explain the bias of clonal lineages to high disease-causing capacity and greater severity or virulence (Coscolla and Gagneux 2014).

Genome-wide sweeps bear the cost of high genetic load (Nielsen 2005), which can be mitigated if the effective population size is large, or by increasing the rate of HGT, which weakens linkage disequilibrium (Parkhill et al. 2003). As neither of these compensatory options is possible in fully clonal pathogens, purifying selection pressure is relaxed and ineffective in removing deleterious mutations. While strongly deleterious mutations, such as those that cause cell death, are removed immediately from a population, the smaller the negative impact on fitness that a mutation has, the slower will be the speed at which it is removed from the population. This time lag is especially marked when the pathogen population is expanding to a new niche, as is the case in a recently emerged or emerging pathogen species. The rate of non-synonymous mutations remains constant relative to the time of divergence while the synonymous mutation rate decreases (Rocha et al. 2006).

The small effective population sizes of specialised pathogens leave them particularly vulnerable to random genetic drift. Transmission events are in effect population bottlenecks, sudden decreases in population size that lead to only a minority of the population surviving to reproduce further, purging diversity and barring variants from future generations. Mutations that increase in prevalence due to genetic drift are the basis for one of the most

widely used methods of pathogen classification, Multi Locus Sequence Typing (MLST). This approach indexes housekeeping genes that are key to the survival of bacteria and form part of the core genome that is typically present in all members of a species. Their sequences can be investigated for neutral (or nearly neutral) mutations that reveal phylogenetic relationships (Maiden et al. 1998; Maiden et al. 2013).

Deleterious mutations may be fixed by such mechanisms in a clonally evolving population, as a consequence of population bottlenecks or through hitchhiking when genome-wide selective sweeps occur. Combined with relaxed purifying selective pressure, deleterious mutations can accumulate in a process known as Muller's Ratchet (Gordo and Charlesworth 2000). During population bottlenecks, small population sizes weaken purifying selection so that random drift becomes the predominant evolutionary force, and deleterious mutations can become fixed by chance. As these mutations are effectively random and can cause genes to lose their function, purifying selective pressure on them is further relaxed, causing gene loss and genome degradation. This has been observed in the emergence of *B. pertussis*, an obligate human pathogen, from the generalist *B. bronchiseptica*, which has a much broader-host specificity and can infect a variety of mammals and birds (Gross et al. 2010). The novel host specificity of *B. pertussis* has been accompanied by and perhaps driven by gene loss. In this case, gene loss followed host restriction, most likely due to the removal of genes that are no longer needed for growth and survival in the environment between host infections (Cummings et al. 2004). This can give the bacteria a fitness advantage by streamlining its genome, removing parts which may have associated costs but no longer incur fitness benefits, but reduction in genome size may instead be a consequence of genome degradation due to Muller's Rachet (Mooi 2010). In the case of *B. pertussis* the reduction in genome size has been accompanied by a massive increase in insertion sequence copy number and a consequent loss of genome structure, which suggest that this has not simply been a streamlining process and that some element of degradative evolution has also occurred (Bart et al. 2014).

A similar situation is observed in *M. tuberculosis*, where gene loss is associated with increased virulence and transmission (Djelouadji et al. 2011). *Mycobacterium leprae*, the agent of leprosy, is an even more extreme example from within the same genus. It is a fully clonal organism, marked by gene decay so extreme that half the genes in its chromosome have become inactive since the bottleneck that formed the emergence of the species (Cole et al. 2001). Homologous recombination, through its origin as a tool for DNA repair, is also an effective way to excise parasitic DNA from the genome. Species that do not participate in HGT are therefore left vulnerable to insertion elements and other forms of parasitic DNA, which may accelerate the degradation of the genome (Croucher et al. 2016). The pattern forms a general syndrome of genome evolution, where host restriction is followed by pseudogene formation, gene loss and increasing numbers of insertion elements (Moran and Plague 2004). This syndrome can even be observed in the parallel evolution of independent clonal lineages. *S. enterica* serovars Paratyphi A and Typhi shared genes by HGT in the recent past, but their subsequent evolution shows evidence of convergence. Both serovars have evolved to a greater degree of host restriction and to cause systemic disease through the loss of function in many of the same genes (Holt et al. 2009).

5. Examples of Adaptive Evolution in Clonal Pathogens

As we have seen, strictly clonal pathogens have limited means by which they can adapt to changes in their environment, being constrained in their ability to generate new variants and to reassort these variants by HGT, even within the same species. Nevertheless, when sufficiently strong, selective pressures can impact the evolution of these pathogens and lead to adaptation based on single mutations that arise in the population. Most often this is observed following positive periodic selection, where a strong selective pressure, following for example a change to the environment, causes a genome-wide sweep that removes any isolates that do not have the allele or alleles most advantageous to the new environment

(Shapiro 2016). In pathogenic bacteria, such events are typically associated with changes caused by medical interventions including the introduction of specific treatments.

B. pertussis has been a significant cause of whooping cough morbidity and mortality in children worldwide, but while the implementation of effective vaccines successfully reduced the disease burden, the bacterium may still be circulating in the population as a milder infection in adults (Guiso 2014). Whole-cell vaccines for *B. pertussis* are highly effective but often cause mild side-effects, and occasionally severe ones. For this reason they were replaced by acellular vaccines in the late 1980s and early 1990s, after which a worldwide selective sweep was observed in the circulating *B. pertussis* strains. The previously dominant *ptxP1* organisms were replaced with strains carrying the *ptxP3* allele, which is associated with an increase in the expression of pertussis toxin. The resulting change in population composition caused phenotypic consequences beyond the pertussis toxin genes, as other mutations hitchhiked with the *ptxP3* variant (de Gouw et al. 2014).

The limited amount of genetic diversity present in clonally evolving populations may belie their adaptive potential. The few existing points of variability in a genetically monomorphic population can have a disproportionally high impact on their biology (Reiling et al. 2013). In *M. tuberculosis*, whole genome sequencing has revealed that the species can adapt when faced with sufficiently strong selection pressure, as is the case with the emergence of antibiotic resistance in the face of treatment. The difficulty in achieving therapeutic doses of antibiotics in the intracellular environment of *M. tuberculosis* compounds the problem of resistance in this species, as therapy with a single antibiotic is rarely effective, and the use of combination therapies accelerates the emergence of multidrug-resistant strains (Moreno-Gamez et al. 2015). Despite the need for each resistance-conferring mutation to occur sequentially within each lineage, resistance is commonplace and multidrug resistance is a growing concern for tuberculosis treatment worldwide (Hershberg et al. 2008).

Hypermutability can enhance the ability of clonal linages to adapt rapidly, and may therefore be actively selected in clonal pathogen populations. This appears to have been the case for the W-Beijing lineage of *M. tuberculosis*, a genotype which is associated with high virulence, in terms of the rate of activation of tuberculosis disease in patients (Merker et al. 2015). This phenotype enhances the spread of tuberculosis in large dense urbanized human populations, but would be less advantageous in less dense human populations where the chronic latent stage, rather than the acute symptomatic stage, would be favored. The mutations promoting virulence in the W-Beijing lineage have probably played a role in the pandemic spread of this lineage during the 20th century (Hanekom et al. 2011). The genetic basis for this increased pathogenicity is variable throughout the lineage, with the common factor appearing to be the relaxation of purifying selection on the genes that control replication and DNA repair, leading to a higher rate of mutations. This increase in genomic variability provides an evolutionary advantage under stressful conditions and can compensate for the loss of adaptability caused by the lack of horizontal gene transfer (Vultos et al. 2008).

S. Typhi has also adapted to the pressure imposed by antibiotic use, with 15 mutations that confer fluoroquinolone resistance being observed at the *gyrA* gene in just a decade. This accumulation of adaptations as a consequence of strong positive selection is in stark contrast to the evidence of neutral evolution in the remainder of *S.* Typhi genome: genetic drift is the predominant force throughout the genome, with no observed difference in the selective forces in genes associated with housekeeping functions or pathogenicity (Roumagnac et al. 2006).

6. Consequences of Clonality

As we have discussed, clonal population structures can lead to specialization and vice versa, leading to the evolution of a highly specialized bacterial clonal pathogen, which has lost

much of its ancestors' metabolic capacity. The ecological and evolutionary consequences of the interaction between clonality and specialisation is a continuous cycle of small effective population sizes that relax purifying pressure and increase the dominance of genetic drift, causing deleterious mutations which may arise to accumulate in the genome. These lead to genome degradation and genetic isolation, which in turn reduces the genetic diversity of a population through sweeps and bottlenecks. Specialising to a specific niche, under these circumstances, can become the only possible evolutionary path that is available to the organism in question, however short-lived that path may ultimately be.

Sexual reproduction is costly for the organisms that engage in it. They must bear the cost of searching and competing for a mate, of producing male offspring, and of sharing only half their genes with any offspring. Why sexual reproduction is widely maintained despite these costs is an "evolutionary puzzle" (Lehtonen et al. 2012). In vertebrates, clonality is always a short-lived strategy that inevitably leads to extinction, though it can be rescued by even occasional recombination (Avise 2015). In bacteria, parasexual processes, mediated by gene transfer and recombination mechanisms, may have evolved primarily as a DNA repair mechanism, belying their far-reaching evolutionary consequences as a method of genetic exchange. The ability of HGT to generate diversity and allow genes, rather than entire genomes, to become the evolutionary units is critical for the long-term success of bacterial species. As with vertebrates, even occasional recombination seems to prevent the complementary processes of clonality and specialisation, by markedly improving the viability of populations over that of a strictly clonal lifestyle (Tibayrenc and Ayala 2012).

When considering the time scales over which clonality can emerge as a stable trait of a given highly specialized pathogen, it is necessary to "distinguish between short term emergence of clonal complexes and the more long-term evolutionary history of a bacterial population" (Feil and Spratt 2001). The concept of clonal lineages emerging from a context with a high frequency of recombination, due to small population sizes that lead to genetic

isolation predates genome sequencing (Levin 1981). Expanding upon this theory led to the epidemic clone model, which posits that even in populations with rampant recombination, successful clones will occasionally cause epidemic waves (Maynard Smith et al. 1993). In this model, a freely recombining population, comprising many variants that participate in HGT, occasionally gives rise to a clonal lineage that successfully emerges as a transmissible pathogen (Turner and Feil 2007).

Are the clonal lineages that survive over millennia fundamentally different from those that survive only until the next genotype in clonal replacement takes over? It is possible that the various pandemics of *Vibrio cholerae*, the cause of the diarrhoeal disease cholera, are an intermediate between short-lived epidemic clones and more long-lived clonal species such as *M. tuberculosis*. Sub-lineages that define each epidemic are observed over the space of a few decades, but are replaced by successive waves of new sub-lineages. These lineages can also develop into hypermutators, again with the pattern of fast adaptive evolution to a specific pressure matched by a decrease in overall fitness (Didelot et al. 2015).

In *Salmonella enterica*, two time-scales of evolution have been proposed, based on studies of the clonal *S.* Typhi serovar. Evolution to antibiotic resistance occurs in clinically observable time, affecting transmission and infections, with a much slower process of genetic drift fixing neutral mutations gradually, resulting in "two distinct epidemiological dynamics" being present in one population structure (Roumagnac et al. 2006). This two-level evolution may explain, at least in part, the predominant nature of clonal evolution being one of stability over thousands of years and marked only by the slow accumulation of mostly neutral mutations, but overlaid with the rapid adaptation in response to the strong selection pressures that are the consequence of antibiotic use or vaccine implementation.

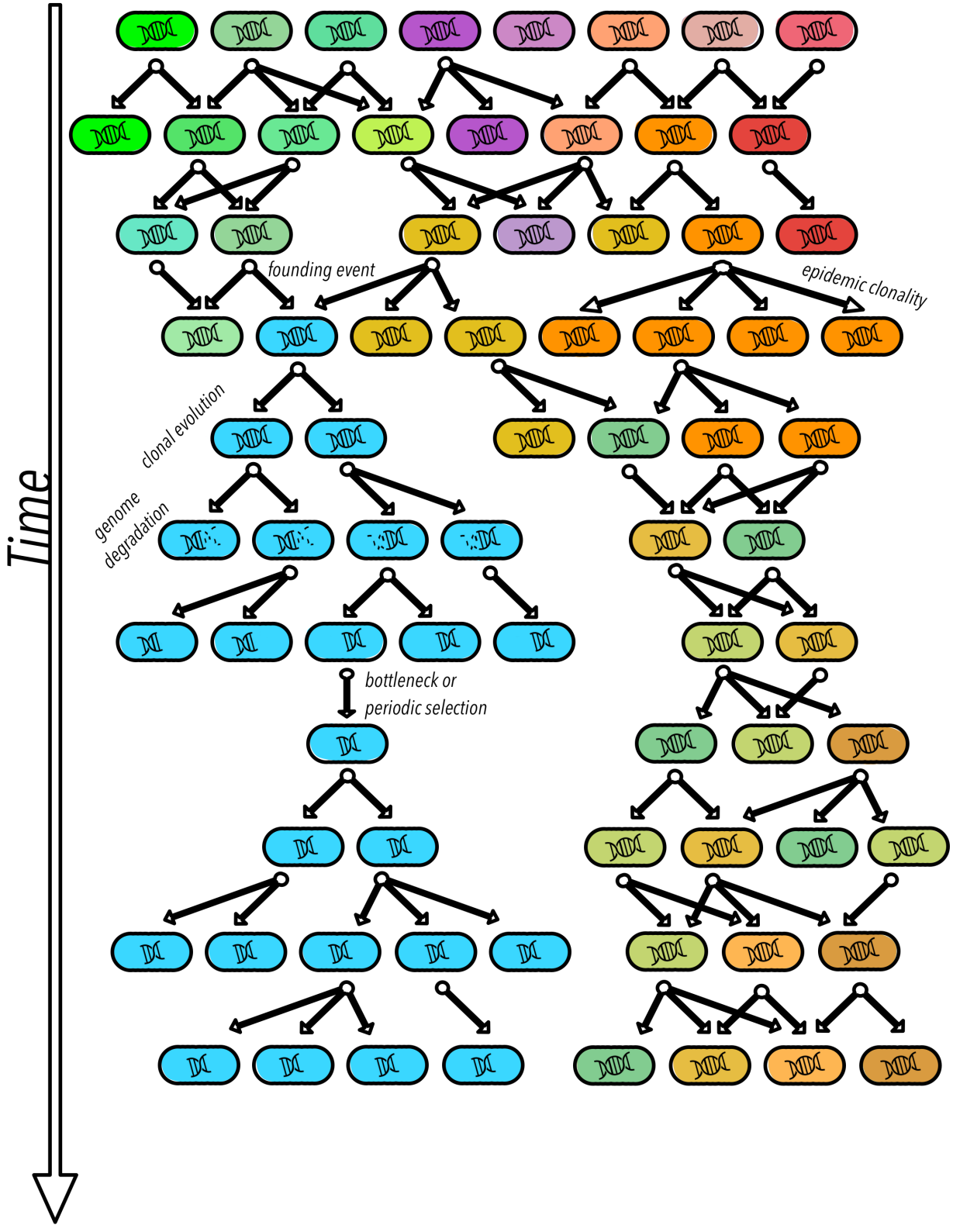
Intriguingly, in another serovar of this species, *S. enterica* serovar Typhimurium, a novel clone appears to currently be emerging as a highly specialised pathogen in observable time. The ST313 sublineage of *S.* Typhimurium causes invasive disease rather than the

gastroenteritis typical of this serovar. Further, the disease that it causes is associated with much higher mortality, with up to half of adult cases resulting in death. The ST313 clone has, through a HGT event, gained a virulence-associated plasmid that bears many antibiotic resistance genes, and this seems to have enabled a change in its ecology that has resulted in a shift in its pathology. The change to an invasive disease niche has subsequently led to genome degradation, including loss of many of the same genes that are now absent from Typhi (Kingsley et al. 2009), in an apparent example of convergent evolution.

Strictly clonal, obligate pathogens of one or a few host species can be seen as a quirk of evolution. They have emerged as specialized clones of large, more adaptable and diverse populations but their invasion of this novel niche, with its consequent process of specialization, comes with the associated costs of loss of adaptability and perhaps leading to a process of degradative evolution that inevitably leads to extinction, even if the host species survives. This process of extinction may take very many host generations and, for a number of human pathogens, this seems to have been ongoing for many millennia, sufficient for global distribution of these seemingly evolutionarily doomed pathogens. Given the high virulence that these pathogens can attain, a fuller understanding of the evolutionary path taken by these bacteria will aid us in containing the long-standing threats that they pose and that may be posed by novel pathogens that take the same evolutionary path as a consequence of changes in human ecology. In this respect, it is worth noting, as a final comment, that our own species is a relative newcomer in terms of worldwide distribution and ecological success (indeed we are unique among the family Hominidae in this respect) and it is therefore unsurprising that the obligate clonal pathogens of humans are also of recent date, with some of them, a notable example being *M. tuberculosis* evolving alongside *Homo sapiens,* with the fate of both species inexorably linked (Soares et al. 2012).

Figure Caption

Figure 1: Illustration of bacterial reproduction and evolution, including epidemic clonality and fully clonal evolution. Each oval represents one bacterium and its genome, and each row represents one sampling time period, so that the bacteria in a row are descendants of those in the row above, separated by many generations. At first the population is evolving as is typical of most bacteria, with a mix of clonal descent and horizontal gene transfer. Therefore most bacteria contain genetic material from more than one ancestor, though not all bacteria contribute genetically to further generations. Between the third and fourth time points, one bacterium is particularly successful and rapidly increases in abundance, in an example of 'epidemic clonality'. This is a temporary trait however, and by the fifth time point onwards the evolutionary processes occurring in its descendants have returned to the patterns typical of the ancestral population, A different bacterium within the population becomes reproductively isolated in the fourth time point, after an HGT event that led to its emergence (the 'founding event'). The descendants of this bacterium undergo 'clonal evolution', with no HGT observable anywhere in the lineage, so that every bacterium is descended from solely one bacterium in a previous time point. This lineage undergoes 'genome degradation' leading to gene loss and eventually to a smaller genome. A 'bottleneck or periodic selection' event dramatically reduces the population size of the clonally evolving lineage between the seventh and eight time points, reducing the diversity of the population and fixing the gene loss in the genome.


Time
founding event
epidemic clonality
clonal evolution
genome degradation
bottleneck or
periodic selection

Chapter I ***References***

Achtman M (2008) Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens. Annu Rev Microbiol 62:53–70. doi: 10.1146/annurev.micro.62.081307.162832

Achtman M (2012) Insights from genomic comparisons of genetically monomorphic bacterial pathogens. Phil Trans R Soc B 367:860–867. doi: 10.1098/rstb.2011.0303

Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6:431–440. doi: 10.1038/nrmicro1872

Avise JC (2015) Evolutionary perspectives on clonal reproduction in vertebrate animals. Proc Natl Acad Sci USA 112:8867–8873. doi: 10.1073/pnas.1501820112

Ayyadurai S, Houhamdi L, Lepidi H, et al (2008) Long-term persistence of virulent Yersinia pestis in soil. Microbiol Immunol 154:2865–2871. doi: 10.1099/mic.0.2007/016154-0

Bapteste E, Boucher Y (2009) Epistemological impacts of horizontal gene transfer on classification in microbiology. Methods Mol Biol 532:55–72. doi: 10.1007/978-1-60327-853-9_4

Bart MJ, Harris SR, Advani A, et al (2014) Global population structure and evolution of Bordetella pertussis and their relationship with vaccination. MBio 5:e01074–e01074–14. doi: 10.1128/mBio.01074-14

Bergstrom CT, McElhany P, Real LA (1999) Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens. Proc Natl Acad Sci USA 96:5095–5100.

Bos KI, Schuenemann VJ, Golding GB, et al (2011) A draft genome of Yersinia pestis from victims of the Black Death. Nature 478:506–510. doi: 10.1038/nature10549

Cao Q, Didelot X, Wu Z, et al (2015) Progressive genomic convergence of two Helicobacter pylori strains during mixed infection of a patient with chronic gastritis. Gut 64:554–561. doi: 10.1136/gutjnl-2014-307345

Cohan FM (2002) Sexual isolation and speciation in bacteria. Genetica 116:359–370. doi: 10.1023/A:1021232409545

Cohan FM (1994) Genetic exchange and evolutionary divergence in prokaryotes. Trends in Ecology & Evolution 9:175–180. doi: 10.1016/0169-5347 (94)90081-7

Cole ST, Eiglmeier K, Parkhill J, et al (2001) Massive gene decay in the leprosy bacillus. Nature 409:1007–1011. doi: 10.1038/35059006

Comas I, Coscolla M, Luo T, et al (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet 45:1176–1182. doi: 10.1038/ng.2744

Coscolla M, Gagneux S (2014) Consequences of genomic diversity in Mycobacterium tuberculosis. Semin Immunol 26:431–444. doi: 10.1016/j.smim.2014.09.012

Croucher NJ, Mostowy R, Wymant C, et al (2016) Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. PLoS Biol 14: e1002394. doi: 10.1371/journal.pbio.1002394

Cummings CA, Brinig MM, Lepp PW (2004) Bordetella species are distinguished by patterns of substantial gene loss and host adaptation. J of Bacteriol doi: 10.1128/JB.186.5.1484

de Gouw D, Diavatopoulos D, Heuvelman K, et al (2014) Differentially Expressed Genes in Bordetella pertussis Strains Belonging to a Lineage Which Recently Spread Globally. PloS One 9:e84523. doi: 10.1371/journal.pone.0084523

Didelot X, Bowden R, Street T, et al (2011) Recombination and population structure in Salmonella enterica. PLoS Genetics 7:e1002191–e1002191. doi: 10.1371/journal.pgen.1002191

Didelot X, Pang B, Zhou Z, et al (2015) The role of China in the global spread of the current cholera pandemic. PLoS Genetics 11:e1005072–e1005072. doi: 10.1371/journal.pgen.1005072

Djelouadji Z, Raoult D, Drancourt M (2011) Palaeogenomics of Mycobacterium tuberculosis: epidemic bursts with a degrading genome. Lancet Infect Dis 11:641–650. doi: 10.1016/S1473-3099 (11)70093-7

Dos Vultos T, Mestre O, Rauzier J, et al (2008) Evolution and diversity of clonal bacteria: the paradigm of Mycobacterium tuberculosis. PloS One 3:e1538–e1538. doi: 10.1371/journal.pone.0001538

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413.

Feil EJ, Spratt BG (2001) Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol 55:561–590. doi: 10.1146/annurev.micro.55.1.561

Ford CB, Lin PL, Chase MR, et al (2011) Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. Nat Genet 43:482–486. doi: 10.1038/ng.811

Gage KL, Kosoy MY (2005) Natural history of plague: Perspectives from more than a century of research. Annu Rev Entomol 50:505–528. doi: 10.1146/annurev.ento.50.071803.130337

Galan JE (2016) Typhoid toxin provides a window into typhoid fever and the biology of Salmonella Typhi. Proc Natl Acad Sci USA 113:6338–6344. doi: 10.1073/pnas.1606335113

Ghodbane R, Mba Medie F, Lepidi H, et al (2014) Long-term survival of tuberculosis complex mycobacteria in soil. Microbiol Rev 160:496–501. doi: 10.1099/mic.0.073379-0

Gordo I, Charlesworth B (2000) The Degeneration of Asexual Haploid Populations and the Speed of Muller's Ratchet. Genetics 154:1379–1387. doi: 10.1016/S0960-9822 (02)00448-7

Gross R, Keidel K, Schmitt K (2010) Resemblance and divergence: the “new” members of the genus Bordetella. Med Microbiol Immunol 199:155–163. doi: 10.1007/s00430-010-0148-z

Guiso N (2014) Bordetella pertussis: why is it still circulating? J Infect 68 Suppl 1:S119–24. doi: 10.1016/j.jinf.2013.09.022

Hanage WP, Fraser C, Spratt BG (2006a) The impact of homologous recombination on the generation of diversity in bacteria. J Theor Biol 239:210–219. doi: 10.1016/j.jtbi.2005.08.035

Hanage WP, Fraser C, Spratt BG (2006b) Sequences, sequence clusters and bacterial species. Phil Trans R Soc B 361:1917–1927. doi: 10.1098/rstb.2006.1917

Hanekom M, Gey Van Pittius NC, McEvoy C, et al (2011) Mycobacterium tuberculosis Beijing genotype: A template for success. Tuberculosis 91:510–523. doi: 10.1016/j.tube.2011.07.005

Hershberg R, Lipatov M, Small PM, et al (2008) High Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human Demography. PLoS Biol 6:e311. doi: 10.1371/journal.pbio.0060311

Holt KE, Parkhill J, Mazzoni CJ, et al (2008) High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nat Genet 40:987–993. doi: 10.1038/ng.195

Holt KE, Teo YY, Li H, et al (2009) Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. Vaccine 25:2074–2075. doi: 10.1093/bioinformatics/btp344

Huttenhower C, Gevers D, Knight R, et al (2012) Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. doi: 10.1038/nature11234

Kay GL, Sergeant MJ, Zhou Z, et al (2015) Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nature Communications 6:6717. doi: 10.1038/ncomms7717

Keim P, Gruendike JM, Klevytska AM, et al (2009) The genome and variation of Bacillus anthracis. Molecular Aspects of Medicine 30:397–405. doi: 10.1016/j.mam.2009.08.005

Kingsley RA, Msefula CL, Thomson NR, et al (2009) Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. Genome Res 19:2279–2287. doi: 10.1101/gr.091017.109

Kohl TA, Diel R, Harmsen D, et al (2014) Whole genome based Mycobacterium tuberculosis surveillance: A standardized, portable and expandable approach. J Clin Microbiol 52:JCM.00567–14–2486. doi: 10.1128/JCM.00567-14

Kolstø A-B, Tourasse NJ, Økstad OA (2009) What sets Bacillus anthracis apart from other Bacillus species? Annu Rev Microbiol 63:451–476. doi: 10.1146/annurev.micro.091208.073255

Lan R, Reeves PR (2001) When does a clone deserve a name? A perspective on bacterial species based on population genetics. Trends Microbiol 9:419–424.

Lehtonen J, Jennions MD, Kokko H (2012) The many costs of sex. Trends in Ecology & Evolution 27:172–178. doi: 10.1016/j.tree.2011.09.016

Levin BR (1981) Periodic selection, infectious gene exchange and the genetic structure of E. coli populations. Genetics 99:1–23.

Maiden MCJ, Bygraves JA, Feil E, et al (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145. doi: 10.1073/pnas.95.6.3140

Maiden MCJ, van Rensburg MJJ, Bray JE, et al (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:728–736. doi: 10.1038/nrmicro3093

Maiden MCJ (2009) Putting leprosy on the map. Nat Genet 41:1264–1266. doi: 10.1038/ng1209-1264

Maynard Smith J, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? Proc Natl Acad Sci USA 90:4384–4388. doi: 10.1902/jop.2011.110063

Merker M, Blin C, Mona S, et al (2015) Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nat Genet 47:242–249. doi: 10.1038/ng.3195

Michod RE, Bernstein H, Nedelcu AM (2008) Adaptive value of sex in microbial pathogens. Infect Genet Evol 8:267–285. doi: 10.1016/j.meegid.2008.01.002

Moffatt CRM, Greig A, Valcanis M, et al (2016) A large outbreak of Campylobacter jejuni infection in a university college caused by chicken liver pâté, Australia, 2013. Epidemiol Infect 144:2971–2978. doi: 10.1017/S0950268816001187

Mooi F (2010) Bordetella pertussis and vaccination: the persistence of a genetically monomorphic pathogen. Infect Genet Evol 10:36–49. doi: 10.1016/j.meegid.2009.10.007

Moran NA, Plague GR (2004) Genomic changes following host restriction in bacteria. Curr Opin Genet Dev 14:627–633. doi: 10.1016/j.gde.2004.09.003

Moran NA, Wernegreen JJ (2000) Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends in Ecology & Evolution 15:321–326.

Moreno-Gamez S, Hilla AL, Rosenbloom DIS, et al (2015) Imperfect drug penetration leads to spatial monotherapy and rapid evolution of multidrug resistance. Proc Natl Acad Sci USA 112:E2874–E2883. doi: 10.1073/pnas.1424184112

Namouchi A, Didelot X, Schöck U, et al (2012) After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. Genome Res 22:721–734. doi: 10.1101/gr.129544.111

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218. doi: 10.1146/annurev.genet.39.073003.112420

Niemann S, Merker M, Kohl T, Supply P (2016) Impact of Genetic Diversity on the Biology of Mycobacterium tuberculosis Complex Strains. Microbiol Spectr. doi: 10.1128/microbiolspec.TBTB2-0022-2016

Ohta T (2011) Near-neutrality, robustness, and epigenetics. Genome Biol Evol 3:1034–1038. doi: 10.1093/gbe/evr012

Ohta T (1992) The Nearly Neutral Theory of Molecular Evolution. Annual Review of Ecology and Systematics 23:263–286. doi: 10.2307/2097289

Park J, Zhang Y, Buboltz AM, et al (2012) Comparative genomics of the classical Bordetella subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. BMC Genomics 13:545–545. doi: 10.1186/1471-2164-13-545

Parkhill J, Sebaihia M, Preston A, et al (2003) Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nat Genet 35:32–40. doi: 10.1038/ng1227

Planet PJ, Narechania A, Chen L, et al (2017) Architecture of a Species: Phylogenomics of Staphylococcus aureus. Trends Microbiol 25:153–166. doi: 10.1016/j.tim.2016.09.009

Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet 29:170–175. doi: 10.1016/j.tig.2012.12.006

Rasko DA, Worsham PL, Abshire TG, et al (2011) Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation. Proc Natl Acad Sci USA 108:5027–5032. doi: 10.1073/pnas.1016657108

Rasmussen S, Allentoft ME, Nielsen K, et al (2015) Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. Cell 163:571–582. doi: 10.1016/j.cell.2015.10.009

Redfield RJ, Findlay WA, Bosse J, et al (2006) Evolution of competence and DNA uptake specificity in the Pasteurellaceae. BMC Evol Biol 6:82. doi: 10.1186/1471-2148-6-82

Reiling N, Homolka S, Walter K, et al (2013) Clade-specific virulence patterns of Mycobacterium tuberculosis complex strains in human primary macrophages and aerogenically infected mice. MBio 4:e00250–13. doi: 10.1128/mBio.00250-13

Rocha E, Maynard Smith J, Hurst LD, et al (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239:226–235. doi: 10.1016/j.jtbi.2005.08.037

Roumagnac P, Weill F-X, Dolecek C, et al (2006) Evolutionary history of Salmonella typhi. Science 314:1301–1304. doi: 10.1126/science.1134933

Shapiro BJ (2016) How clonal are bacteria over time? Current Opinion in Microbiology 31:116–123. doi: 10.1016/j.mib.2016.03.013

Smith NH, Gordon SV, la Rua-Domenech de R, et al (2006) Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat Rev Microbiol 4:670–681. doi: 10.1038/nrmicro1472

Soares P, Alshamali F, Pereira JB, et al (2012) The Expansion of mtDNA Haplogroup L3 within and out of Africa. Mol Biol Evol 29:915–927. doi: 10.1093/molbev/msr245

Spratt BG, Maiden MCJ (1999) Bacterial population genetics, evolution and epidemiology. Phil Trans R Soc B 354:701–710. doi: 10.1098/rstb.1999.0423

Stackebrandt E, Frederiksen W, Garrity GM, et al (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043–1047. doi: 10.1099/00207713-52-3-1043

Suerbaum S, Smith JM, Bapumia K, et al (1998) Free recombination within Helicobacter pylori. Proc Natl Acad Sci USA 95:12619–12624.

Tibayrenc M, Ayala FJ (2002) The clonal theory of parasitic protozoa: 12 years on. Trends Parasitol 18:405–410.

Tibayrenc M, Ayala FJ (2015) How clonal are Neisseria species? The epidemic clonality model revisited. Proc Natl Acad Sci USA 112:8909–8913. doi: 10.1073/pnas.1502900112

Tibayrenc M, Ayala FJ (2016) Is Predominant Clonal Evolution a Common Evolutionary Adaptation to Parasitism in Pathogenic Parasitic Protozoa, Fungi, Bacteria, and Viruses? Elsevier Ltd

Tibayrenc M, Ayala FJ (2012) Reproductive clonality of pathogens: A perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. Proc Natl Acad Sci USA 109:E3305–E3313. doi: 10.1073/pnas.1212452109

Turner KME, Feil EJ (2007) The secret life of the multilocus sequence type. Int J Antimicrob Agents 29:129–135.

van der Veen S, Tang CM (2015) The BER necessities: the repair of DNA damage in human-adapted bacterial pathogens. Nat Rev Microbiol 13:83–94. doi: 10.1038/nrmicro3391

Votintseva AA, Miller RR, Fung R, et al (2014) Multiple-Strain Colonization in Nasal Carriers of Staphylococcus aureus. J Clin Microbiol 52:1192–1200. doi: 10.1128/JCM.03254-13

Wiedenbeck J, Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35:957–976. doi: 10.1111/j.1574-6976.2011.00292.x

Woodford N, Ellington MJ (2007) The emergence of antibiotic resistance by mutation. 13:5–18. doi: 10.1111/j.1469-0691.2006.01492.x

Yahara K, Didelot X, Jolley KA, et al (2016) The Landscape of Realized Homologous Recombination in Pathogenic Bacteria. Mol Biol Evol 33:456–471. doi: 10.1093/molbev/msv237

Chapter II
GUAVA: Rapid analysis of sequence diversity and selection in gene-by-gene annotated data sets

Sofia Hauck, James E. Bray and Martin C. J. Maiden

Department of Zoology, University of Oxford, United Kingdom OX1 3PS

Abstract

Background: Gene-by-gene annotation is an efficient and scalable method for genome analysis, particularly well suited to the analysis of thousands of genomes simultaneously. Multi-locus sequence typing, abbreviated to MLST, uses this method. With the advent of high-throughput sequencing, MLST can be expanded to around fifty genes in the universal ribosomal MLST (rMLST) scheme or to genus- or species-specific schemes, which can index sets including thousands of loci. Many tools exist to extract phylogenetic information from databases that use gene-by-gene allele annotations to describe isolates according to these MLST schemes, but there is a need for tools that can use this data to examine population-level traits such as quantifiable signals of diversity and selection.

Results: GUAVA uses allele designations as the basis for a rapid and comprehensive quantification of diversity and selection strength at each annotated locus. Loci are evaluated for their length and number of unique sequences before and after translation. Allele-based measures analogous to single nucleotide polymorphism density and to the ratio of

synonymous to non-synonymous mutations are calculated from these values, as the number of unique alleles per base pair length and the ratio of unique sequences pre- and post-translation. GC content, and details about the number of allele designations are also reported. The results are presented in a universal text format, which can be viewed in a visualisation tool that is available for local and online use. This tool includes adjustable tolerance for data quality, sort-and-filter dynamic data tables, and interactive graphical plot generation and labelling.

Conclusion: The method presented here allows for computational efficiency and ease of analysis of gene diversity data within large genome collections. This allows for comprehensive studies of genetic diversity for thousands of isolates simultaneously, even when repeated analysis is required, for example for routine surveillance or in resource-limited settings. It is designed to have few software dependencies, and to be easy to use for users unfamiliar with bioinformatics software.

Availability: The GUAVA application is comprised of two parts, a processing phase as a Perl script and a visualisation phase as an R script. Both can be downloaded from the supplemental information of this article, or from https://github.com/sohauck/GUAVA.The visualisation step can also be accessed directly as a website at https://shauck-zoo.shinyapps.io/GUAVA/.

Keywords:

bacterial population genomics	genetic diversity
whole genome sequencing	cgMLST
gene-by-gene analysis	wgMLS

Chapter II: ***Table of Contents***

Abstract **1**

Keywords: **2**

Background **4**

List of abbreviations **9**

Results & Discussion **10**

Table 1. List of the fifteen loci with highest *AllelicDiv* value 12

Figure 1. Distributions of parameters calculated for 3,993 loci in a dataset of 7,656 Mycobacterium tuberculosis isolates. 15

Speed of Processing **16**

Figure 2. Change in runtime (in CPU seconds) as the number of isolates or the number of loci included in the analysis increases. 17

Visualisation **17**

Figure 3. Screenshot of GUAVA's visualisation interface. 20

Accessibility **21**

Conclusion **22**

Methods **23**

Test Data Set (*M. tuberculosis*) **23**

Processing alleles **24**

Variables in output **24**

Validation **25**

Figure 4. Correlation between allele-based and alignment-based measures. 27

Visualisation **27**

Portability **29**

Declarations & Supplementary files **30**

References **31**

Background

Whole genome sequencing (WGS) is becoming an increasingly routine procedure in microbiology as costs rapidly decrease (Kyrpides et al. 2014). WGS can provide high-resolution discrimination of isolates, allowing it to function as the ultimate typing method for bacteria (Pallen et al. 2010). Because of this combination of high resolution and low cost, many systems of epidemiological surveillance are founded on WGS technology, leading to the creation and maintenance of databases with up to tens of thousands of genomes sampled from individual bacterial populations (Brisse et al. 2014; Kohl et al. 2014). However, there is a need for tools that analyse these massive data sets for traits observed at a population level, and especially for tools that can accomplish this when utilised by non-experts, as WGS has many applications outside of academic research (Salipante et al. 2015). Such tools must minimise repetition and maximise speed, in order to tolerate the continued addition of new records to genome databases and to ensure they can be deployed in real-time for routine analyses or in resource-limited contexts. Here we develop such a tool, using gene-by-gene annotations to investigate the genetic diversity, signatures of selective pressure and other traits of genes in bacterial populations.

Due to the inherent portability and scalability of the gene-by-gene approach, it can be used to annotate genomes in large databases. The most common application of this approach is known as multi-locus sequence typing (MLST) (Maiden et al. 1998). The first MLST schemes were species- or genus-specific and consisted of six or seven loci, typically gene fragments that would then be sequenced directly with the use of specific oligonucleotide primers. With the advent of WGS, there now exists a universal bacterial typing scheme that utilises 53 ribosome protein subunits (ribosomal MLST or rMLST) (Jolley et al. 2012) as well as core genome or whole genome MLST (cgMLST or wgMLST). In these population annotation approaches, two complementary databases interact to catalogue loci, identifiable regions of genomes. The first database holds isolate records, including sequence data (in the

case of assembled draft genomes, as a set of contigs), and any metadata, such as when and where the isolate was sampled. The second, a nomenclature database, holds reference sequences for the loci to be examined, including all known alleles of every locus, each of which has a unique allele identifier. Typing is applied to these populations to characterise them into clusters following this population annotation. When MLST was first introduced, a scheme could have as few as six loci, but in wgMLST this can be up to thousands, each with hundreds of known alleles, and these numbers are only expected to increase as more data are collected. By combining the two databases, the genome of each isolate can be automatically searched, by BLAST or another sequence searching algorithm, for any matches to the alleles present in the nomenclature definitions database. This process is called "scanning." Matches, if they are exact, can be recorded immediately, adding a new allele designation at that locus for that isolate, in what is termed "tagging." Non-exact matches can be submitted as potential novel alleles to the nomenclature database, so that "scanning and tagging" is an iterative process that becomes more exhaustive as new data is added. As more sequences are added to the sequence definitions database, annotations of genomes become more complete, and as more genome sequences are added, more alleles become known (Maiden et al. 2013).

While loci can be any definable region of a genome, including coding and non-coding fragments, here we use complete coding sequences, which make up the majority of bacterial genomes (Mira et al. 2001). These sequences are readily identified by their start and stop regions, and each nucleotide sequence can be translated into a sequence of amino acids that may form the basis of a protein. These proteins carry out the majority of all cell functions, including metabolism, signalling and reproduction, and the function of a protein, by the means of its impact on the fitness of the organism, affects the evolution of its associated locus. By determining the location and precise sequence of loci, automated 'scanning' enables the gene-by-gene approach to represent an entire genome as a profile of loci designations with allele identifiers. Conceptualising genes as discrete units, found in the genome as loci and capable of modifying the biochemistry of an organism through its proteins, allows for an

intuitive understanding of genome biology. For example, the acquisition of a toxin gene or to a mutation that creates a novel antibiotic resistance allele are both examples of a genomic change in the form of a single gene alteration with a direct effect on the biology of the organism (Maiden and Harrison 2016). Like any model, this is a simplification that misses some of the nuance of reality — functional units may be made up of multiple genes that have no phenotypic impact in isolation, for example.

Choosing the loci of interest and their reference sequences, and curating the genome annotations created by the scanning process, are specialist tasks that require knowledge of the genetics of the organism in question. Scanning genomes is a resource-intensive process. Increasing levels of automation have lowered the human time required for curation at the cost of increasing the computational resources required, but some manual curation is still required, especially when databases are first established. The stability of annotations ensures that this does not need to be repeated for previous records as new isolates or new loci are added, but only in the rare instance that a locus definition is changed. Downstream analyses therefore have good reproducibility and speed without losing the capacity to improve with new information (Bratcher et al. 2014). Additionally, the runtime for such analyses scales linearly as data is added and results are portable as long as the nomenclature database is shared or harmonised, so that allele numbers refer to the same allelic sequences across multiple databases (Maiden and Jolley 2010). Due to these benefits in scalability and portability, many collections now numbering in the hundreds or thousands of whole genome sequences are managed with gene-by-gene curation. Allele designations have so far been used largely for investigating the phylogenetic relatedness of isolates (Jolley and Maiden 2013), but their link to protein function means they also carry information about the biology and evolution of organisms. A few MLST-like schemes have been designed specifically to investigate a set of phenotypic traits, such as comparing disease isolates with vaccine types in the Bexsero Antigen Sequence Typing (BAST) scheme (Brehony et al, 2016). In these cases

the loci under investigation are few in number, for example only five loci in the BAST scheme. In contrast, the wgMLST scheme used here for *M. tuberculosis* has nearly four thousand loci.

The gene-by-gene approach is also well suited to analysing data that includes gaps. Short-read sequences, of 200 base pairs or fewer each, are common outputs for high-throughput sequencing technologies that have precipitated the boom in WGS. Genomes assembled from short-read sequences using de novo assembly may have repetitive sequences or gaps in coverage that make it impossible to confidently place all reads, so that the genome may be represented not by one long uninterrupted sequence but a set of contiguous sequences (contigs) (Bratcher et al. 2014). Alleles can be discovered in any of these contigs, as the "scanning" process looks for matches to known alleles through all contigs. A central advantage of the reference-free assembly process is that previously undiscovered genes and genes that are not present in the reference are nevertheless assembled. In the case of gene-by-gene annotation, this may mean that novel genes can be discovered and found in previously assembled genomes, without the need to re-assemble or re-map reads. This is an important characteristic when investigating species without closed genomes, and species where the pan genome, the complete catalog of all genes in that species, and the genome of any individual isolate, differ markedly in size.

Genome collections that represent a comprehensive sampling of one population allow for the examination of the diversity of the "gene pool" existing within it, covering both the majority of the pan genome and the majority of alleles for known loci. Genetic mutations arise randomly with regard to gene function, such that in a population under no selective evolutionary pressure all loci exhibit similar degrees of sequence diversity, observed as an even distribution of alleles. Even mutations with small effects on fitness will not have their frequency affected, as sufficiently strong genetic drift can overpower signals of selection, making them impossible to detect (Ohta 1992). Therefore, any significant deviations from the genome's background level of diversity are notable indicators of selection. For example, a

locus exhibiting markedly low diversity may be under purifying pressure, perhaps because its encoded protein is essential for the survival of the bacteria (Jolley et al. 2012). Likewise, high diversity could point to an immunological function that maintains co-occurring alleles within a single population (Buckee et al. 2007). Evidence of strong diversifying selection, observed as an excess of non-synonymous mutations when compared to synonymous mutations that do not change the amino acid chain of a protein implicates that locus as one that is adapting to a changing environment (Kryazhimskiy and Plotkin 2008). This knowledge, beyond further characterising the evolutionary path of a population, can inform development of vaccines, drugs (Agüero et al. 2008), and diagnostic tests (Jolley et al. 2007).

GUAVA is the Gene-by-gene Unified Analysis and Visualising Application. It aims to make the population-level traits about individual loci easier to observe by addressing the current lack of scalable tools that quantify signals of diversity and selection for very large datasets of whole genomes. Because it completes its analysis based on discrete allele designations, not alignments, it is fast, processing thousands of genomes with thousands of loci in minutes, and it is scalable, with total duration time that rises linearly with added samples. Due to its speed and compatibility to existing notations, it has the potential to expand the knowledge that can be acquired from existing data sets, and to become a routine part of surveillance work in the future.

List of abbreviations

- ***WGS:*** Whole genome sequencing
- ***cgMLST:*** Core genome multi-locus sequence typing
- ***wgMLST:*** Whole genome multi-locus sequence typing

GUAVA output variables:

- ***Locus:*** The name of locus as it appears in the allele designation table.
- ***Missing:*** The number of isolates for which no allele designation was present in the table.
- ***MultipleCopies:*** The number of isolates in which the locus had at least two tagged alleles, *usually in the case of paralogous genes.*
- ***CountNuc:*** The count of unique nucleotide sequences for specified locus.
- ***CountAA:*** The count of unique amino acid sequences for specified locus.
- ***MinLength:*** The length in nucleotide letters (base pairs) of shortest nucleotide sequence.
- ***MaxLength:*** The length in nucleotide letters (base pairs) of longest nucleotide sequence.
- ***AvgLength:*** The average length in nucleotide letters (base pairs) of unique nucleotide sequences.
- ***GC content:*** The proportion of G or C base pairs in all alleles (as opposed to A or T).
- ***AllelicDiv:*** The number of unique nucleotide sequences found divided by average length. Formula: *CountNuc / AvgLength*
- ***RatioCount:*** The ratio of unique nucleotide to unique amino acid sequences. Formula: *CountAA / CountNuc*
- ***AliLenNuc:*** The length of the alignment of all sequences in this locus in nucleotide format.
- ***AliLenAA:*** The length of the alignment of all sequences in this locus in amino acid format.
- ***VSitesNuc:*** The proportion of sites in the alignment of nucleotide sequences which had any variation.
- ***VSitesAA:*** The proportion of sites in the alignment of amino acid sequences which had any variation.
- ***RatioVS:*** The ratio of proportions of unique sites in nucleotide to amino acid sequences.

Results & Discussion

GUAVA was used to evaluate the genetic diversity of 3,993 loci in a set of 7,656 *Mycobacterium tuberculosis* isolates, as shown in Figure 1. The genetic variability in *M. tuberculosis* is low compared to other bacterial species, with a maximum of around 1,800 single nucleotide polymorphisms separating any two human strains (Coscolla and Gagneux 2014). The isolate collection used here was chosen to reflect the global diversity of *M. tuberculosis*, with representatives from all of its major lineages. These lineages are phylogeographically structured and are known to vary in their pathobiology, including their transmissibility and virulence across diverse human ethnic groups. Because the species is highly clonal, with no trace of horizontal gene transfer, descent is unambiguous and the genome is closed, so that all genes are present in all isolates. While GUAVA can accommodate both recombination and accessory genomes, these lack of these features in *M. tuberculosis* simplify the analysis, making this an ideal model species in which to demonstrate and validate the application of this method.

For each of the 3,993 loci, the total number of unique allele sequences and the length of these sequences were measured. Dividing the former by the latter gives a measure named of allelic diversity named *AllelicDiv*, for short, which represents the number of alleles observed per length in base pairs. This is an allele-based metric that is analogous to the sequence-based metric of single nucleotide polymorphism (SNP) density, with the key difference that any mutation (not only SNPs but also insertions or deletions) is included as a polymorphism. The values of *AllelicDiv* in this data set were found to have a distribution that closely clustered around the genome mean (0.061 alleles per nucleotide), with 86.7% (3,465 loci) within one standard deviation (0.0222) from it, with outliers at extremely high values up to a maximum over 20 standard deviations above the mean (0.523 alleles per nucleotide). The strong clustering around the genome average is consistent with the previous studies on the genetic diversity within *M. tuberculosis*. Because it is a non-recombining, or clonally evolving,

highly specialised pathogen, its effective population size is very small. Small founding populations are strong affected by the hitchhiking effect and by reduced purifying pressure, so that *M. tuberculosis* evolution occurs largely through drift, fixing even slightly deleterious mutations (Hirsh et al. 2004). Out of the 15 loci exhibiting the highest levels of diversity as measured by *AllelicDiv*, also listed in Table 1, seven were categorised as cell wall components, including five ESAT-6-like proteins. These proteins are antigens that provoke a strong immune response in patients with tuberculosis (Brandt et al. 2000), and may be co-adapting with specific human populations (Comas et al. 2013).

GUAVA also calculates a value that resembles the ratio of synonymous to non-synonymous mutations, named *RatioCount*, by taking the ratio of unique sequences before and after translation. It is analogous to the sequence-based Ka/Ks ratio (often known as the dN/dS ratio) (Kryazhimskiy and Plotkin 2008), and indicates the strength and direction of selective pressures acting on a locus. The maximum value of this measure is 1, where all sequences have unique translations. Therefore, in datasets where the genetic distances between isolates are great, alleles may be sufficiently distant that they each have a unique translation, causing *RatioCount* to be saturated and leading to a loss of resolution. This is especially likely to happen in loci with short alleles or in smaller datasets where the total count of alleles is low.

Table 1. List of the fifteen loci with highest *AllelicDiv* value

Locus name	Product	Unique alleles	Average length (in base pairs)	*AllelicDiv* value
MYCO000095	Unknown	215	411	0.5231
MYCO003919	Gid (streptomycin resistance)	274	675	0.4059
MYCO002043	PncA (pyrazinamide resistance).	207	560.9	0.369
MYCO002346	EsxO (putative ESAT-6 like protein)	97	285	0.3404
MYCO003478	PPE60	356	1180	0.3017
MYCO001793	EsxN (putative ESAT-6 like protein)	81	285	0.2842
MYCO000397	Unknown	101	365.6	0.2763
MYCO003862	WhiB6 (regulates ESX-1 transcription)	96	351	0.2735
MYCO000277	Possible toxin VapC25	110	429	0.2564
MYCO000010	Unknown	92	419.4	0.2194
MYCO001197	EsxK (putative ESAT-6 like protein)	51	297	0.1717
MYCO003648	CspA (probable cold shock protein A)	35	204	0.1716
MYCO001037	EsxI (putative ESAT-6 like protein)	47	285	0.1649
MYCO000749	Possible toxin VapC31	68	429	0.1585
MYCO003619	EsxV (ESAT-6 like protein)	42	285	0.1474

While alignments are omitted by default, due to their slow calculation speed, they can be requested if a nucleotide-level analysis is needed. Two parameters, *VSitesNuc* and *VSitesAA* are calculated from the data, representing the proportion of sites with variability in a multiple sequence alignment, in both the nucleotide and amino acid formats, respectively. The ratio of these can be taken, giving a measure of the ratio of polymorphic site densities (*RatioVS*), somewhat analogous to Ka/Ks ratio, as both refer to the rate of non-synonymous to synonymous mutations. Similarly to the earlier *RatioCount*, these values are vulnerable to saturation in diverse data sets. In data sets that include samples that are very distant from one another, every site may have variability, so the difference in diversity between genes is not distinguishable by this method. Alignment-based measures are less susceptible to this

saturation than allele-based ones, as they consider each alignment site independently, while any mutation in an allele is enough to mark it as entirely distinct.

While the results from the *GUAVA* analysis of this data were compared against sequence-based methods, these methods are not fully equivalent and it not expected that they would give identical results, even in clonally evolving populations. Multiple polymorphic sites, each counted and weighed independently in sequence-based methods, are condensed into a single allelic difference in allele-based methods. For example, a mutation causing a ten-codon removal is evaluated as a change with a 30-fold greater magnitude in a sequence-based as opposed to an allele-based method, despite both being caused by a single mutation event.

The gene-by-gene approach simplifies sequences into designations early in the process, causing a loss of resolution which limits the power of analyses (Didelot and Falush 2007). This disadvantage is mitigated by the capacity to use all polymorphisms, with no need to filter SNPs and no discovery bias. Since allele identifiers can be used to refer back to nucleotide sequences, further investigations can recover all individual polymorphisms for later analyses at a higher level of detail if necessary. Therefore, allele-based methods are particularly useful for preliminary or large-scale analyses, such as routine surveillance or investigation into the nature of large databases.

The stability of allele designations, especially important in large or growing data sets, is another advantage of allele-based methods, granting both faster calculations in subsequent analyses and reproducibility even when more data is added. This is true especially for datasets comprised of closely related isolates, for example belonging to the same species, or even the same sub-species cluster.

Allele-based measures such as *AllelicDiv* and *RatioCount*, as compared to their sequence-based analogous measures, have an advantage in speed of calculation and a disadvantage in a

loss of resolution. They also differ in measuring genetic differences by either their relative impact on the genotype (sequence-based, where every changed nucleotide is counted separately) or by an approximation of the number of HGT events or mutations (allele-based, where differing alleles are counted).

In the case of *RatioCount*, this difference in measuring alleles as opposed to nucleotides in a sequence also limits the scope of mutations covered by this metric. It is well-suited for quantifying the ratio of synonymous to non-synonymous mutations in populations undergoing primarily purifying selection, but it is not suitable to populations undergoing other primarily other kinds of selection, or for populations whose members are not closely related (due to the saturation of the metric when all alleles represent unique amino acid sequences).

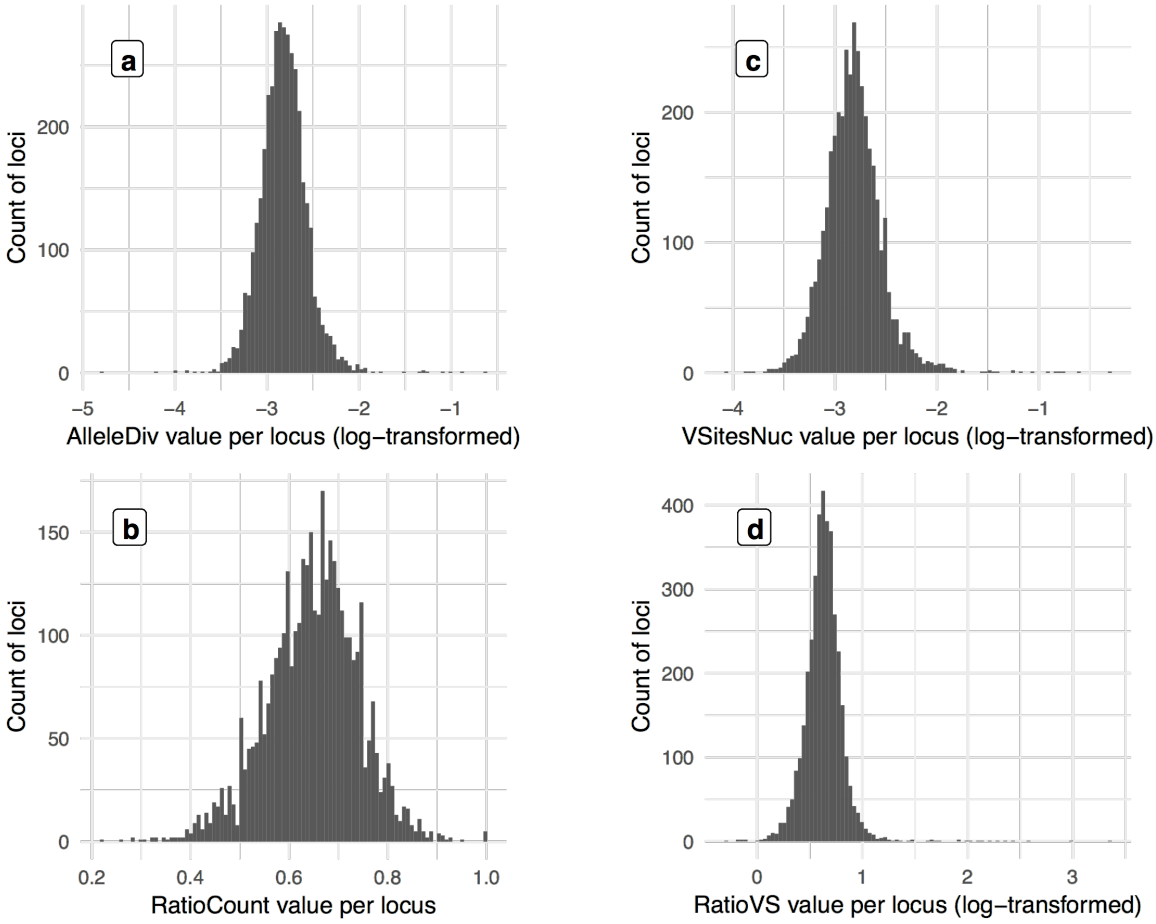


Figure 1. Distributions of parameters calculated for 3,993 loci in a dataset of 7,656 Mycobacterium tuberculosis isolates.

a. The distribution of the logarithm of the allele-based measure of genetic diversity, *AllelicDiv*, defined as the number of alleles found per average allele length in base pairs. **b.** The distribution of the allele-based measure of selection strength, *RatioCount*, defined as the count of unique amino acid sequences divided by the count of unique nucleotide sequences. **c.** The distribution of the logarithm of the sequence-based measure of genetic diversity, *VSitesNuc*, defined as proportion of sites with variations in a multiple sequence alignment of nucleotide format alleles. **d.** The distribution of the sequence-based measure of selection strength, *RatioCount*, defined as the ratio of the proportion of variable sites in multiple sequence alignments in amino acid and nucleotide formats.

The values measured by GUAVA are robust to various sources of error. Because loci are considered as independent units within a genome and all alleles of one locus are pooled before analysis, results are not distorted by horizontal gene transfer or by mixed samples (as long as both the main and contaminant samples are within the study population). Single base-call errors, the most common type of sequencing error outside of homopolymeric repeat regions (Metzker 2010), can be hard to detect, as they may appear indistinguishable from single point mutations. To control for this, alleles may be filtered by observed frequency. For example, including only alleles seen in at least two genomes takes advantage of the low probability that the same non-systematic sequencing error will occur in two samples. Due to the low frequency of most alleles, this option is recommended only for large data sets.

Validation of both *AllelicDiv* and *RatioCount* against their respective analogous sequence-based measures was carried out and is described in detail in the Methods of this paper. It showed that *AllelicDiv*, can appear to underestimate diversity in loci with multiple nucleotides differing, as expected when contrasting allele-based to sequence-based methods. *RatioCount* had more limitations, as a poorly predictive of selection strength both as it approaches 1 (due to saturation, where all alleles have unique amino acid sequences), and where its sequence-based analogue, Ka/Ks, was low, due to low resolution present in a collection of few alleles.

Speed of Processing

GUAVA processed the test data set, comprised of 3,993 loci and 7,656 isolates, with over 30 million pre-calculated allele designations, in less than four minutes on an off-the-shelf desktop computer with a 2.7GHz Intel Core i5 processor and 8 GB RAM.

The total runtime increases linearly with the addition of more loci, and roughly linearly with the addition of new genomes, as can be observed in Figure 2. Due to the uneven distribution of alleles, with some having high frequencies and most being observed only

rarely, a small number of genomes may contain a majority of all alleles in a population. Because the processing time increases most directly in relation to the number of alleles under review, an analysis containing a relatively small number of genomes may have a slightly longer than expected processing time. However, when most alleles have been discovered, the addition of more isolates causes a less-than-proportional increase in the processing time.

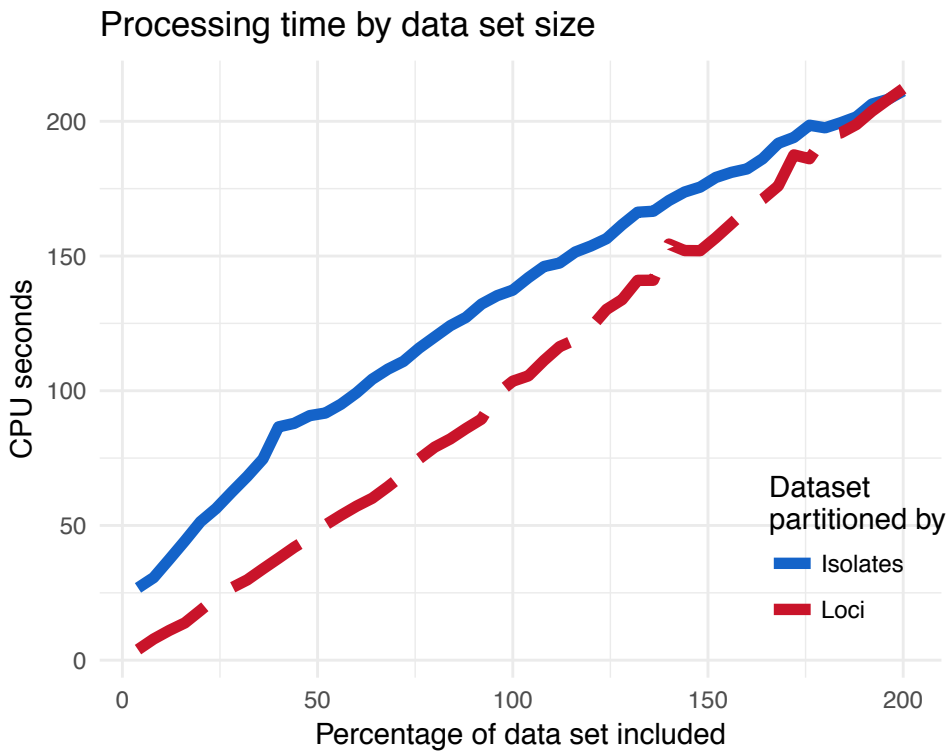


Figure 2. Change in runtime (in CPU seconds) as the number of isolates or the number of loci included in the analysis increases.

Visualisation

The visualisation tool, built with the Shiny (Chang et al. 2015) package of R (RDCT 2011), allows for results of the analysis to be explored intuitively and comprehensively. It uses a point-and-click interface that requires no knowledge of coding, and offers a download option for exporting the figures created, as is shown in Figure 3. The output table from GUAVA's processing phase can be uploaded to the visualiser through a graphical file selection window

that opens following a click to the “Browse…” button. Once data has been loaded, it instantly appears as a scatter plot where each point represents one locus. To the right of that plot is the “Control Panel”. Any changes in the left-hand control panel take effect immediately on the main graph display.

Loci with few allele designations may appear less genetically due to a limited set of observations. Excluding loci based on how often they have been observed can be done using a slider on the Control Panel. Loci which have allele designations in fewer isolates than the minimum percentage chosen in the slider are automatically excluded from the analysis. Returning the slide to a lower level restores the excluded loci. Directly below the slider, the Control Panel displays the total number of isolates reviewed and the number of loci currently in the analysis, as well as the total possible number of loci, the current minimum designation level, and the number of loci filtered out by this level as well as the number manually removed. By clicking on a tab at the top of the main graph display, a set of graphs specifically designed to aid in choosing a minimum designation percentage cut-off can be accessed. They display the distribution of the number of isolates for which each locus has no allele designation, and the relationship between how often a locus has no designation and its estimated genetic diversity.

In the lower half of the Control Panel, “Plotting Options” can be chosen. Options include choosing which parameters are plotted, in both the positions of the points and their colour. The horizontal axis can be simply randomised, alphabetical by locus name, or according to locus category if this was included in the input table. The vertical axis can plot *AllelicDiv* and *RatioCount* values in all cases, as well as *VSitesNuc*, *VSitesAA* and *RatioVS* if alignments were completed. The average across all loci is plotted as a dashed blue line for each of these varibles. Point colours can reflect the number of isolates for which a locus has no allele designations, or for which the locus has multiple allele designations, as well as the average length of its alleles or a user-inputted category. A legend for the meaning of the point colours

is displayed below the graph, and dynamically updates as necessary, as do the axes labels. Two checkboxes also give the option to use z-scores instead of the true values in the labels of the vertical axis, or to use the logarithm of values, which also activates Bonferroni-corrected significance thresholds for outlier points utilising the p-value of 0.05, as dotted green lines.

Another user-friendly feature of the visualizer is the ability to label specific points. An area of the current main plot can be selected by dragging over it, which creates a highlighted overlay. This overlay can be repositioned for better accuracy, or cancelled by a single click outside it. Double-clicking the overlay marks the loci plotted in the area that the overlay covers as "labelled" loci. This causes these loci to be plotted in a different colour to unlabeled loci (as long as no other variable has been chosen for the point colours). It also causes the name of the locus to be displayed above its plotted point. This is especially useful for identifying outliers. Since this label selection remains in place as any parameters in the visualiser are changed, specific loci can be tracked over various metrics using this feature. Alternatively, this feature allows for specific loci to be chosen prior to their manual removal from the plot. By pressing the "Remove labelled points" button, any loci currently in the "labelled" status are manually removed from the analysis. This can be reversed by a click to the "Reset Removals" button.

A third tab provides a table with the data itself, with options to search, filter, and sort dynamically by any of the variables in the table. For example, it is possible to sort the table by which points have been labelled, allowing for a quick look up of previously discovered outliers, or the search for a specific locus name. A "Download table" button allows the table to be saved locally as a text file with a single-click.

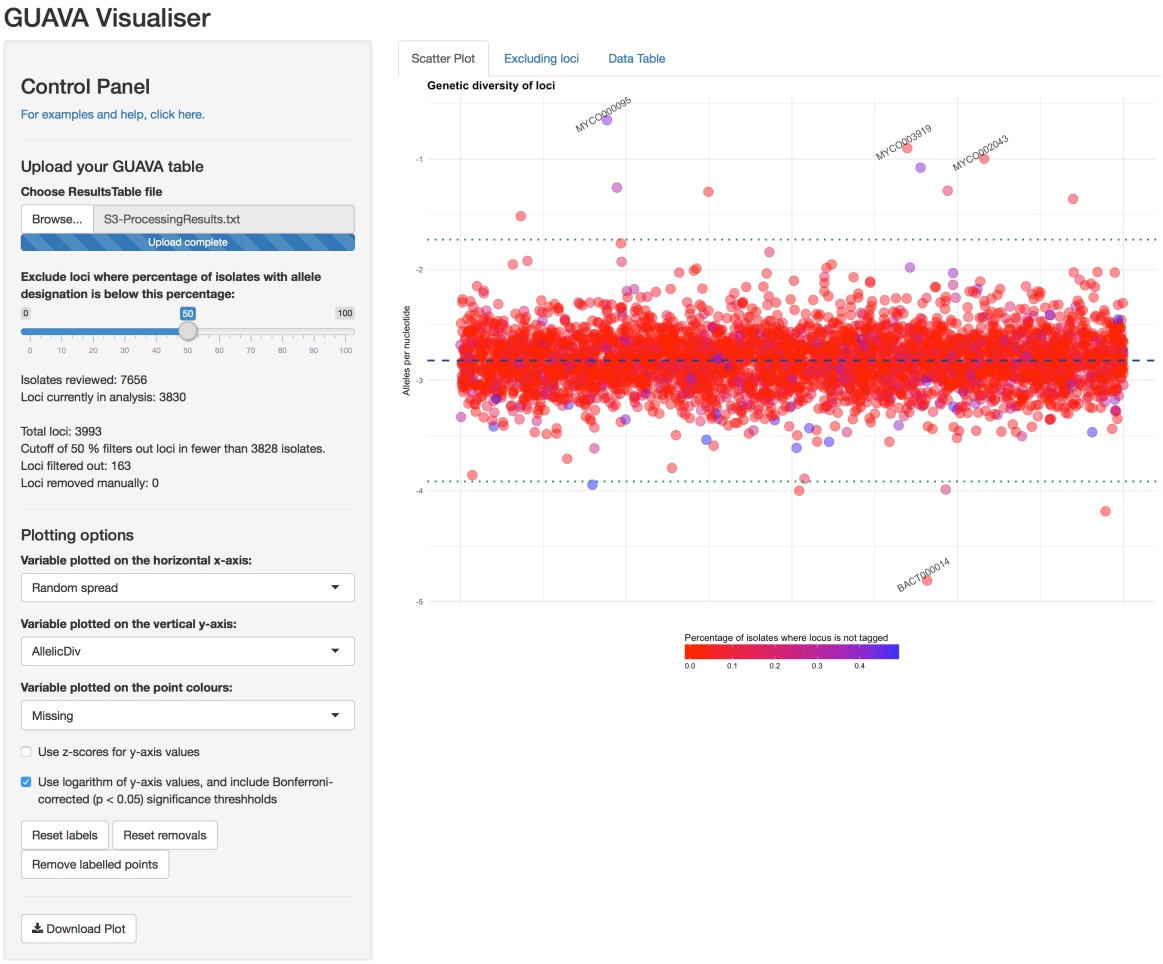


Figure 3. Screenshot of GUAVA's visualisation interface.

The left-hand column shows the control panel, which includes, from top to bottom: a link with examples and instructions, a button to open the file selection window to upload data, a slider to choose how strictly to consider loci for inclusion in analysis and a few lines explaining the effects of the chosen slider value, three drop-down menus to choose parameters to plot points in location (horizontal and vertical) and colour, checkboxes with options to label with z-scores, use logarithm values and significance thresholds, and finally buttons to remove labelled points, reset labels and removals, and download the plot. The main area shows the current plot, in this case, the distribution of the *AllelicDiv* value, a measure of genetic diversity, with the mean value marked by dashed blue line, and the percentage of missing allele designations in the colour of points, where red is always present and blue is missing in half of isolates. Four loci are labeled, with their locus name appearing

in the main plot near their plotted points. Tabs along the top show the other main area options, namely two additional plots to clarify the effect of the slider for locus inclusion, and a dynamic data table that may also be downloaded.

Accessibility

GUAVA does not need to be installed before use, but has a handful of dependencies that are free to install and use. Perl is required in the processing phase (tested on version 5.18.2), while MAFFT (tested on version 7.221) is required only if the optional alignments are requested. While this script must be run on the command line and options can be added directly on the command line, the script can detect missing inputs and query for them. For example, if no address is given for an output folder, a prompt will ask for it, or allow the option to be left blank to use the default, which is a new directory in the same location as the input allele designation table.

The visualisation tool can be run locally if R (tested on version 3.3.2) and its Shiny package (tested on version 1.0.4) are installed, or accessed freely through the website: https://shauck-zoo.shinyapps.io/GUAVA/. The visualisation tool is a point-and-click interface, with options for modifying plots in real life using pull-down menus, excluding or labelling specific loci, and downloading plots and tables directly.

The main input required for *GUAVA* is a table of allele designations, in which loci and isolates form the rows and columns, or vice versa, and cells contain the unique allele identifiers that refer to specific allele sequences. Tables like as these are a common way to express allele designation profiles in various gene-by-gene annotation software packages, such as BIGSdb (Maiden and Jolley 2010), SeqSphere (Juenemann et al. 2013) and BioNumerics (www.applied-maths.com/bionumerics). It is also a common format for visualisation software, such as PhyloViz (Nascimento et al. 2017). The main output is also a tab-separated table file, which may be read by any text or spreadsheet editor and has a small file size (for example, 383 kilobytes in the case of the 3,993 loci used in this test dataset),

making them ideal for sharing data and for storing results in a portable format. All sequences are obtained and saved in the universal FASTA format, with the exception of alignments which are in the CLUSTALW format. Both can be read by any text-editor.

Conclusion

GUAVA uses existing expert curation to rapidly evaluate signals of sequence diversity and selection pressure in collections of thousands of genomes with thousands of loci. Each locus is independently measured for its values across a number of variables. The distribution of these values, especially its outliers, can reveal the non-random patterns which indicate evolutionary pressures that act on specific loci. Using genome annotations directly for these calculations boosts the utility of existing curated databases, and allows for rapid calculation. GUAVA includes both a powerful processing phase and an included visualisation tool (available online or locally). Results from the processing phase can be easily shared and explored in the visualisation tool, with no knowledge of coding required to create figures or sort and filter data. This allows for a rapid overview of the diversity of loci across a genome even for large data sets, combining a comprehensive but fast analysis. Furthermore, the inclusion of an intuitive built-in visualisation streamlines the process of hypothesis generation and provides a straightforward method for hypothesis testing or for examining the nature of genetic diversity at a population level.

Methods

The GUAVA application consists of two parts: a processing phase written in Perl (version 5.18.2), which optionally uses MAFFT (version v7.221) for multiple sequence alignments, and a visualisation tool which utilises an R (version 3.3.2) Shiny (version 1.0.4) interface for data exploration. A test data set of 7,656 *M. tuberculosis* genomes, annotated with reference to 3,993 genes present in finished H37Rv genome, was used to perform runtime and validation tests.

Test Data Set (M. tuberculosis)

The data set used for the validation of this method was a collection of 7, 656 *Mycobacterium tuberculosis* genomes added to the PubMLST database and analysed using BIGSdb platform, using its automated genome annotation features.

The loci analysed were genes described in the H37Rv (NC_000962.3) reference genome (Camus et al. 2002). Only protein-coding loci with complete coding sequences were included, with RNA genes, genes with frameshifts or internal stop codon, and genes missing either start codons or stop codons excluded. The allele sequences of each locus in the H37Rv reference were used as "seed sequences" for the annotation. That is, the alleles present in H37Rv were used as the reference from which other alleles were defined, based on their sequence identity and alignment length.

A table of allele designations is the main data input format for GUAVA. A table transposition option is included in the processing phase, so that tables in either isolates-as-rows or isolates-as-columns format can be used.

Multiple allele designations for one locus in a single isolate can be indicated by cells that contain multiple allele IDs separated by semi-colons. This is most frequently the case for paralogous loci. Paralogues can be the result of gene duplications, and cause multiple genes in a genome to fit within the sequence identity and alignment length criteria that define

alleles in a locus. It can also occur when genome assemblies were computed from short read sequences belonging to multiple isolates, causing two alleles to exist at one locus.

Processing alleles

The frequency of all alleles in one locus, across all genomes in the dataset, is calculated from the input table of allele designations. For every allele that appears above the user-determined frequency cutoff threshold (the default is 1, for all observed alleles), the associated nucleotide sequence is acquired from the sequence source. The sequence source can be a local directory containing the sequences in FASTA format (the fastest option), or the name of an open access PubMLST nomenclature database. If a database name is given as the source, it is possible to copy the database's entire collection of alleles for the included loci into a local directory, so this copy can be used as a local source in the future. Alternatively, it is possible to access the required sequences only, directly from the database, but this is the slowest method per allele.

Variables in output

The variables in the output of GUAVA are summarized in Page 9, in the List of Abbreviations. They are fifteen in total, excluding the locus name, which is listed used to identify loci. At each defined locus, the total count of unique nucleotide sequences (*CountNuc*) is recorded. The counts of the number of isolates at which the locus is found to be have multiple allele designations (*MultipleCopies*), and at which there is missing data, with no allele designation given (*Missing*) are also noted. The minimum, maximum and average lengths of the alleles, in base pairs, are measured (*MinLength*, *MaxLength* and *AvgLength*). The proportion of all base pairs which are G or C, as opposed to A or T, is also measured (*GCContent*). All unique sequences are translated, and a count of unique translated sequences present is noted (*CountAA*). Both "raw" and translated allele pools can be aligned using the MAFFT program (Katoh and Standley 2013) (arguments for which can be set by the user, or left to run under its default settings). The alignments can then be evaluated to give the total

number of positions without variation and the total alignment length (recorded as AliLenNuc and AliLenAA for the before and after translation alignments respectively).

Five measures of diversity and selection are calculated from these values. The first of these quantifies genetic diversity and is named *AllelicDiv*. This is simply *CountNuc* divided by *AvgLength*, or that is, the number of unique sequences per locus divided by that locus's average allele length. The count of unique sequences before and after translation can be divided. This value, named *RatioCount*, simply *CountAA* divided by *CountNuc*, gives an estimate of the ratio of synonymous to non-synonymous mutations.

If alignments were requested, the number of variable sites can be divided by the alignment length to obtain the proportion of sites with variation, which is recorded as *VSitesNuc* and *VSitesAA* for the measure taken from the nucleotide and amino acid alignments respectively. The ratio of these measures can be taken, as *VSitesAA* divided by *VSitesNuc*, giving a single measure of the ratio of polymorphic sites density (*RatioVS*).

Validation

The alleles of each locus were "reverse translated" using *TranslatorX* (Abascal et al. 2010) in combination with MAFFT. The SeqinR package of R was used to obtain Ks and Ka matrices for the alignments of each locus (Charif and Lobry 2007). Sequence-based measures of diversity and selection were obtained by adding Ks, the number of substitutions per synonymous site between protein-coding genes, to Ka, the equivalent for non-synonymous sites (Hurst 2002) and by diving Ka by Ks, respectively.

The *AllelicDiv* measure was contrasted to the sum of Ks and Ka matrices for validation. The correlation between these values (both transformed using the logarithm function) was highly statistically significant ($p < 0.0001$), with a Pearson's product moment correlation coefficient estimated at 0.8544 (95% CI: 0.8456–0.8628). A simple general linear model similarly found a coefficient of determination of 0.730. *AllelicDiv* measured a lower value

relative to the sum of Ks and Ka for some loci, as represented as points above the blue line in Figure 4a. This is an expected effect because as an allele-based metric, *AllelicDiv* measures only whether alleles differ at all, not in how many locations of their alignment they differ. It therefore can appear to underestimate diversity, as wildly different alleles are counted as having the same magnitude of difference as two alleles that vary at a single nucleotide.

The *RatioCount* value was contrasted to its equivalent metric, the Ka/Ks ratio. The correlation between these values (following a log-transformation of Ka/Ks) was highly statistically significant ($p < 0.0001$), with a Pearson's product moment correlation coefficient estimated at 0.841 (95% CI: 0.831–0.850). A simple general linear model found a coefficient of determination (adjusted R-squared value) of 0.707. *RatioCount* generally correlates well with Ka/Ks, with two exceptions: first, as *RatioCount* approaches 1 (where every allele has a unique amino acid sequence) it can be an underestimate of the strength of purifying selection when compared to Ka/Ks, and secondly, *RatioCount* can be an overestimate compared to Ka/Ks for loci with a very low Ka/Ks ratio, as shown in Figure 4b as points below the blue line. In this data set, these loci have only a few alleles (due mainly to being short loci with few nucleotides), giving them poor resolution and therefore poor statistical power. A large data set may be able to mitigate this.

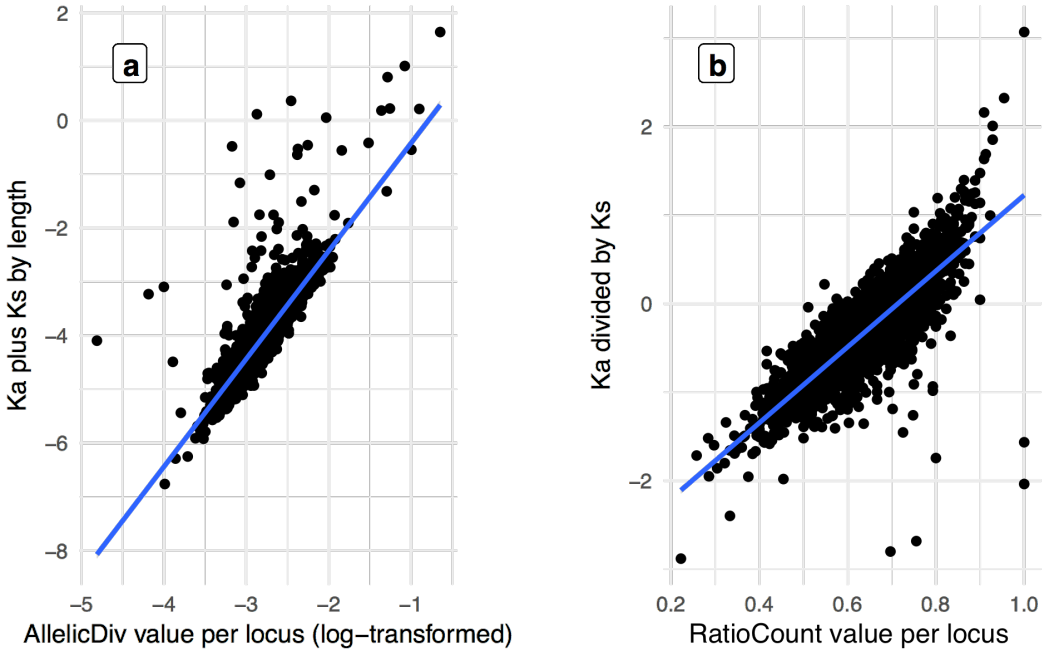


Figure 4. Correlation between allele-based and alignment-based measures. **4a.** The correlation at each locus between the log transformed values of *AllelicDiv* and the sum of Ka and Ks, the number of non-synonymous and synonymous mutations in a multiple sequence alignment of all alleles at that locus, divided by the nucleotide length, for all 3,993 loci. **4b.** The correlation at each locus between *RatioCount* value and the Ka (the number of non-synonymous mutations) divided by Ks (the number of synonymous mutations) over a multiple sequence alignment of all alleles per each of 3,993 loci.

Visualisation

Visualisation of the results is included in *GUAVA*, via an R Shiny application that can be run locally or accessed online. The application has a point-and-click interface that is accessed via an Internet browser window. The left-hand side column is the Control Panel and the main panel on the right displays plots and tables. Inputs by the user, including the table of variables given by the processing phase of GUAVA, are handled using reactive programming so that the outputs (plots, tables and panels) are dynamically and instantly updated in response to any changed to the inputs by the user.

The Control Panel includes a link to the GitHub page with examples, a "read me" file, and a license file. Below it is a file-selection button that opens a file selection browser window. Selecting a file immediately uploads it and presents a plot on the right with the default plot options. Further options in the Control Panel change the main plot. A slider from with values 0-100% allows for the exclusion of loci based on their "Missing" value, i.e. the number of isolates in which the loci have no designation. Loci with a "Missing" value that are greater than the percentage chosen on the slider multiplied by the number of isolates are excluded from analysis. An information panel immediately below the slides gives the count of isolates in the original analysis, the count of loci currently remaining in the analysis, the total number of loci included, and how many loci according to this criterion or excluded manually.

The "Plotting Options" in the Control Panel include three drop-down menus, two checkboxes and three buttons. The drop-down menus select the x-axis variable, the y-axis variable and point colours variables. The options available for the user in each of these drop-down menus changes dynamically in accordance with the variables in inputted table. For example, only if alignments were requested in the processing phase will the proportion of variable sites appear as an y-axis plotting option. The check-boxes allow the y-axis to be labelled by z-scores, and for the use of the logarithm-transformed values for the y-axis. The latter also triggers the plotting of two threshold lines, for the values at which points are significantly different from the mean value of all loci, given a p-value of 0.05 and including a Bonferroni multiple-testing correction. These thresholds appear as dotted green lines while the mean across all loci is plotted as a blue dashed line. Buttons at the bottom of the "Plotting Options" allow for labels to be reset, removals to be reset and for the manual removal of points which are currently labelled. Below "Plotting Options" is a button that allows for the download of the plot (as a PDF file) or the table, changing dynamically according to which tab is chosen.

The main plot is visible in the first of three tabs. It can be modified, by choosing the variable plotted on either axis and in the point colours, as well as changing the axes labels to z-scores, or using log-transformed values. By brushing over the graph, points can be selected, and with a double-click, labeled with the name of the loci they represent. This label will remain even if plot options are changed, so points can be "followed" through various data plotting options. The status of loci as labelled or unlabelled is added as a variable to the data table, so points can be sorted or filtered by their label status. By clicking the "Remove labelled points" button, loci currently labelled can be removed from the analysis, including the plot and table, until "Reset removals" is pressed. All of these changes occur in real time, so users can immediately see the effect of their choices on the main plot display. The plot can at any time be downloaded (as a PDF file) with a single click.

The third tab displays a dynamic data table. Each of the columns can be used to sort the table, or filter it, by typing search terms. This data table includes only the loci within the tolerance set by the slider, and an added column that marks the loci labeled via the brush-and-click mechanism. The table can be downloaded (as a tab-separated CSV file) in a single click.

Portability

All of these variables are recorded by GUAVA's processing phase into a single table, with a new row for each examined locus. The table's header also includes the software version used, the number of genomes reviewed and the date at which it was created. This table is a small (383 kilobytes in this test data set), text-only file that can easily be stored in a portable format, read in by any spreadsheet or text editor, and easily shared. The sequences obtained from the sequence source, their translations, and if requested, alignments in both nucleotide and amino acid format are all saved as well in the results directory, but they are not necessary for the integrity of the main results table and can be deleted if they are not useful.

Declarations

- Authors' contributions: SH designed and completed the scripting, analysis and manuscript; JEB performed the genome assemblies and gave advice on Perl scripts; JEB and MCJM provided advice on manuscript.
- Availability of data and material: The datasets generated and/or analysed during the current study are available in PubMLST (https://www.pubmlst.org). The code used in available at GitHub (https://github.com/sohauck/GUAVA).

Supplementary files

S1. Allele Designation Table for 7,656 M. tuberculosis isolates over 3,993 genes. TXT file, 60.5 MB in size.

S2. Locus information table, with functional categories and gene names. TXT file, 107 KB in size.

S3. GUAVA Processing phase results. TXT file, 383 KB in size.

S4. KaKs values table. TXT file, 993 KB in size.

S5. Speed benchmark table. 835 bytes in size.

Chapter II ***References***

RDCT (2011) R: A Language and Environment for Statistical Computing. The R Foundation for Statistical Computing, Vienna, Austria

Abascal F, Zardoya R, Telford MJ (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res 38:W7–W13. doi: 10.1093/nar/gkq291

Agüero F, Al-Lazikani B, Aslett M, et al (2008) Genomic-scale prioritization of drug targets: the TDR Targets database. Nature Reviews Drug Discovery 7:900–907. doi: 10.1038/nrd2684

Brandt L, Elhay M, Rosenkrands I, et al (2000) ESAT-6 subunit vaccination against Mycobacterium tuberculosis. Infect Immun 68:791–795.

Bratcher HB, Corton C, Jolley KA, et al (2014) A gene-by-gene population genomics platform: de novo assembly, annotation and genealogical analysis of 108 representative Neisseria meningitidis genomes. BMC Genomics 15:1138. doi: 10.1186/1471-2164-15-1138

Brehony C, Rodrigues CMC, Borrow R, et al (2016) Distribution of Bexsero® Antigen Sequence Types (BASTs) in invasive meningococcal disease isolates: Implications for immunisation. Vaccine 34:4690–4697. doi: 10.1016/j.vaccine.2016.08.015

Brisse S, Brehony C, Conceição T, et al (2014) Microbial molecular markers and epidemiological surveillance in the era of high throughput sequencing: An update from the IMMEM-10 conference. Research in Microbiology 165:140–153. doi: 10.1016/j.resmic.2014.01.003

Buckee C, Danon L, Gupta S (2007) Host Community Structure and the Maintenance of Pathogen Diversity. Proceedings: Biological Sciences 274:1715–1721. doi: 10.2307/25249238

Camus JC, Pryor MJ, Medigue C, Cole ST (2002) Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiol Immunol 148:2967–2973.

Chang W, Cheng J, Allaire J, et al (2015) Shiny: web application framework for R.

Charif D, Lobry JR (2007) SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M (eds) Structural Approaches to Sequence Evolution. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 207–232

Comas I, Coscolla M, Luo T, et al (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet 45:1176–1182. doi: 10.1038/ng.2744

Coscolla M, Gagneux S (2014) Consequences of genomic diversity in Mycobacterium tuberculosis. Semin Immunol 26:431–444. doi: 10.1016/j.smim.2014.09.012

Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–1266. doi: 10.1534/genetics.106.063305

Hirsh AE, Tsolaki AG, DeRiemer K, et al (2004) Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proc Natl Acad Sci USA 101:4871–4876. doi: 10.1073/pnas.0305627101

Hurst LD (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 18:486–487. doi: 10.1016/S0168-9525 (02)02722-1

Jolley KA, Bennett JS, Maiden MCJ, et al (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiol 158:1005–1015. doi: 10.1099/mic.0.055459-0

Jolley KA, Brehony C, Maiden MCJ (2007) Molecular typing of meningococci: recommendations for target choice and nomenclature. FEMS Microbiol Rev 31:89–96. doi: 10.1111/j.1574-6976.2006.00057.x

Jolley KA, Maiden MCJ (2013) Automated extraction of typing information for bacterial pathogens from whole genome sequence data: Neisseria meningitidis as an exemplar. Euro Surveill 18:20379–.

Juenemann S, Sedlazeck FJ, Prior K, et al (2013) Updating benchtop sequencing performance comparison. Nat Biotechnol 31:294–296. doi: 10.1038/nbt.2522

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. doi: 10.1093/molbev/mst010

Kohl TA, Diel R, Harmsen D, et al (2014) Whole genome based Mycobacterium tuberculosis surveillance: A standardized, portable and expandable approach. J Clin Microbiol 52:JCM.00567–14–2486. doi: 10.1128/JCM.00567-14

Kryazhimskiy S, Plotkin JB (2008) The Population Genetics of dN/dS. PLoS Genetics. doi: 10.1371/journal.pgen.1000304

Kyrpides NC, Hugenholtz P, Eisen JA, et al (2014) Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains. PLoS Biol. doi: 10.1371/journal.pbio.1001920

Maiden M, Bygraves JA, Feil E, et al (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145. doi: 10.1073/pnas.95.6.3140

Maiden MCJ, Harrison OB (2016) The population and functional genomics of the Neisseria revealed with gene-by-gene approaches. J Clin Microbiol 54:JCM.00301–16–1955. doi: 10.1128/JCM.00301-16

Maiden MCJ, Jolley KA (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595. doi: 10.1186/1471-2105-11-595

Maiden MCJ, van Rensburg MJJ, Bray JE, et al (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:728–736. doi: 10.1038/nrmicro3093

Metzker ML (2010) Sequencing technologies — the next generation. Nat Rev Genet 11:31–46. doi: 10.1038/nrg2626

Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589–596.

Nascimento M, Sousa A, Ramirez M, et al (2017) PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics 33:128–129. doi: 10.1093/bioinformatics/btw582

Niemann S, Merker M, Kohl T, Supply P (2016) Impact of Genetic Diversity on the Biology of Mycobacterium tuberculosis Complex Strains. doi: 10.1128/microbiolspec.TBTB2-0022-2016

Ohta T (1992) The Nearly Neutral Theory of Molecular Evolution. Annual Review of Ecology and Systematics 23:263–286. doi: 10.2307/2097289

Pallen MJ, Penn CW, Loman NJ (2010) High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Current Opinion in Microbiology 13:625–631. doi: 10.1016/j.mib.2010.08.003

Salipante SJ, SenGupta DJ, Cummings LA, et al (2015) Application of Whole-Genome Sequencing for Bacterial Strain Typing in Molecular Epidemiology. J Clin Microbiol 53:1072–1079. doi: 10.1128/JCM.03385-14

Chapter III
Lineages of Mycobacterium tuberculosis differ in their patterns of genetic diversity

Sofia Hauck[1], James E. Bray[1], Keith A. Jolley[1] and Martin C. J. Maiden[1]

[1]*Department of Zoology, University of Oxford, United Kingdom OX1 3PS*

Abstract

Background: Whole genome sequencing is the only technique that provides enough resolution for the accurate characterisation of *Mycobacterium tuberculosis* species complex (MTBC) isolates. Therefore, large collections of MTBC whole genome sequence isolates exist, from which the population genetics of the species complex can now be explored, including the genetic changes driving the higher virulence observed in Lineage 2 and Lineage 4.

Methods: A core genome scheme for the *Mycobacterium* genus was developed using the finished genomes of 19 species in the genus. The 749 loci in this scheme were used to assign a collection of 6,679 MTBC isolates to six lineages within *M. tuberculosis* or to other species in the complex. 3,795 loci were assessed in this isolate collection for their genetic diversity and selection using both allele-based and alignment-based methods. These measures were taken for all lineages, and for Lineage 2 and Lineage 4 in isolation. Homoplasy between Lineage 2 and Lineage 4 was also assessed.

Results: We find that the *Mycobacterium* core genome scheme provides sufficient resolution to assign the majority of MTBC isolate into distinct sub-species lineages. The genetic diversity of loci was highly variable, with an over-representation of virulence and drug resistance associated genes displaying the highest diversity. Genes associated with the ESAT-6 family, and genes involved in iron uptake, stress response, antibiotic resistance were the drivers of adaptation in Lineage 2 and Lineage 4.

Conclusions: This analysis presents an overview of the diversity within the MTBC. It finds many loci previously highlighted by experimental means for their important function in the MTBC have genetic patterns that differ from the genome at large. However, many loci with unknown functions appear to be evolving distinctly from the rest of genome. Further work into the function of these unknown genes is necessary for a fuller understanding of the MTBC.

Table of Contents

Abstract **1**

Background **6**

Methods **10**

Loci **10**

wgMLST: whole genome MLST 10

rMLST: ribosomal MLST 11

cgMLST: core genome MLST 11

Table 1. Finished genomes used in *Mycobacterium* genus core scheme creation pool 12

Total loci in analysis 12

Table 2. Loci used in this analysis and their functional categories according to the H37Rv annotation, and their abbreviations. 13

Isolates 13

Table 3. Species and genome status of isolates included in analysis. 14

Genome assembly 14

Table 4. The minimum, maximum and mean values for assembly statistics of the draft genomes. 14

Figure 1. The relationship between contig number and N50 contig length for assembled draft genomes. 15

Genome Annotation 15

Lineage assignment 16

Phylogenetic tree 17

Measures of diversity, selection and homoplasy 18

Results 21

Curation of isolates and loci 21

Figure 2. Distribution of loci according to the number of isolates in which they had no allele designations. 22

Lineage assignment 22

Table 5. Groups in the cgMLST classification scheme with a threshold of 120 allele mismatches. 24

Table 6. Groups in the classification scheme with a threshold of 75 allele mismatches, found within Group 1 of the classification scheme with a threshold of 120 allele mismatches. 24

Table 7. Lineage assignments for all isolates in which this was possible. 25

Phylogenetic structure 25

Figure 3. A Neighbour-Joining phylogenetic tree of 6,661 isolates in the MTBC. 26

Figure 4. Pairwise distances between 6,679 MTBC isolates, across 749 loci in the *Mycobacterium* genus cgMLST scheme. 27

Figure 5. Pairwise distances within the major lineages of the MTBC. Pairwise comparisons of five or fewer allele mismatches were removed. 28

Table 8. Pairwise distances within the major lineages of the MTBC and in total for the species complex. 28

Signals of diversity and selection 29

Table 9. Loci with allelic diversity values significantly different from genome average. 30

Figure 6. Allelic diversity values, log-transformed for 3,795 loci in the MTBC. 31

Table 10. Loci with allele ratio values significantly different from genome average. . 32

Figure 7. Allele ratio values for 3,795 loci in the MTBC. 33

Figure 8. Correlation between Griffin et al (2016) essentiality p-value and the allele ratio. 34

Table 11. Loci with proportions of variable sites in a nucleotide multiple sequence alignment that differ significantly from genome average. 36

Figure 9. Proportions of variable sites in a nucleotide multiple sequence alignment, log-transformed for 3,795 loci in the MTBC. 37

Table 12. Loci with proportions of variable sites in an amino acid multiple sequence alignment that differ significantly from genome average. 38

Figure 10. Proportions of variable sites in an amino acid multiple sequence alignment, log-transformed for 3,795 loci in the MTBC. 39

Table 13. Loci with a ratio of proportions that differ significantly from genome average. 41

Figure 11. Ratio of proportions of variable sites in protein to nucleotide multiple sequence alignments, log-transformed for 3,795 loci in the MTBC. 42

Table 14. Loci significantly different values of allelic diversity between Lineage 2 and other lineages. 43

Table 15. Loci significantly different allele ratio values between Lineage 2 and other lineages. 44

Table 16. Loci significantly different values of allelic diversity between Lineage 4 and other lineages 45

Table 17. Loci significantly different allele ratio values between Lineage 4 and other lineages. 46

Homoplasy 47

Table 18. Loci with high levels of homoplasy, as determined by a high proportion of alleles shared between Lineage 2 and Lineage 4. 48

Discussion 49

Allele designations 49

Lineage assignment and phylogenetic structure 49

Diversity and selection 51

Lineage-specific effects 53

Lineage 2 53

Lineage 4 54

Further work 55

Supplemental files 56

References 57

Background

The *Mycobacterium tuberculosis* species complex (MTBC) consists of closely related species, all of which can cause the disease tuberculosis. These include human-adapted lineages, namely *M. tuberculosis* sensu stricto (Lineages 1-4) and *M. africanum* (Lineage 5 and 6); several species adapted to other mammals, such as *M. bovis* (cattle), *M. microti* (voles), *M. caprae* (sheep and goats), *M. orygis* (antelope), *M. mungi* (mongoose), and *M. pinninpedii* (seals and sea lions); and lastly a recently discovered species with a uncertain reservoir (sampled from a human host), *M. canettii* (Gagneux 2013).

Tuberculosis has plagued humans for millennia, and causes a significant disease burden still today, with 10.4 million new cases and 1.4 million deaths estimated worldwide in 2015. While the rate of both new cases and deaths is falling, an end to the epidemic remains a distant possibility (World Health Organization 2016). Up to a third of the world's population is infected by the MTBC, but only one in ten succumb to the disease. Host factors such as age, malnutrition, immune status and stress are the main factors shaping the clinical outcome of tuberculosis patients (Donoghue 2009). Combined with the difficulty of assessing genetic differences between MTBC strains due to their low sequence diversity (Comas et al. 2009), this led to a "prevailing dogma" in which genetic variation between human-adapted strains of the MTBC was considered either non-existent or clinically irrelevant (Gagneux and Small 2007).

Studies made possible by advances in whole genome sequencing have upended both aspects of this dogma. While the species complex is still considered to harbour little variation in comparison to other bacterial species, and this small degree of diversity is relatively likely to affect the pathobiology of the organism due to an unusually high proportion of non-synonymous mutations (Hershberg et al. 2008). While not every non-synonymous mutation is guaranteed to cause a change in the pathobiology of the organism, they are far more likely to do so than

synonymous changes that do not affect protein structure but may affect transcription rates. A variety of genetic typing methods exist for the identification of MTBC strains, but they all find a congruent phylogeny of species, lineages and sub-lineages (Niemann et al. 2016).

While *M. tuberculosis* is found worldwide, the distribution of its lineages around the globe is not uniform. All species within the MTBC, with the possible exception of *M. canettii*, are obligate pathogens, and the majority are specialised to infect humans. This co-existence, combined with the emergence of the MTBC in parallel to *Homo sapiens*, means the geographic distribution of the pathogen lineages reflects the evolutionary path they share with humans (Comas et al. 2013). Beyond the phylogenetic mirroring of population structure that emerges as the consequence of a common history, functional co-adaptation results from this host-pathogen relationship (Gagneux 2012).

Among the strongest evidence of this co-adaptation is the fact that transmission rates are higher when the host and pathogen are sympatric. In the case of *M. tuberculosis*, this refers to cases where the place of origin of human host and the place of origin of the MTBC lineage match one another. The genetic basis for this can be confirmed by how the sympatric effect can be observed in global cities such as San Francisco, where both host and pathogen are away from their place of origin and disparate human ethnic groups and tuberculosis lineages exist in close proximity (Hirsh et al. 2004). The intense level of co-adaptation, in light of the limited levels of genetic variation in the human and MTBC populations, is further highlighted by the impact of AIDS on this relationship. Co-infection with HIV erases the effect of sympatry. That is, in the presence of HIV, the weakened immune system of the human host is equally vulnerable to all human-adapted lineages of *M. tuberculosis* (Fenner et al. 2013). This pinpoints the main feature of this co-adaptation as the interaction between the human immune system and the proteins expressed by *M. tuberculosis* cells.

Although six decades have passed since studies were published demonstrating lineage-specific differences in virulence (Bhatia et al. 1961), the genetic determinants of virulence in *M. tuberculosis* remain unclear. Efforts to find such determinants were hindered by the fact that unlike many pathogenic species, *M. tuberculosis* has no "canonical" virulence factors, and by the lack of genetic typing methods with sufficiently high resolution to characterise variation between strains until recently (Coscolla and Gagneux 2014).

An exception to this lack of clarity exists for genetic changes linked to drug resistance. Multi-drug resistant (MDR) tuberculosis is a growing threat, with significant under-reporting and under-treatment of cases. The rise of extensively drug resistant (XDR) tuberculosis further complicates control of this epidemic, with strains that are resistant to both first-line and second-line drugs (Gandhi et al. 2010; World Health Organization 2016). Dozens of mutations related to the resistance against drugs used to treat tuberculosis have been discovered, including mutations that directly confer resistance by changing either protein structures or intergenic promoter regions, and compensatory mutations that mitigate the fitness costs of the resistance-conferring mutations (Cohen et al. 2015).

In this study, we investigate genetic diversity in the MTBC, with a special focus on Lineage 2 and Lineage 4 of *M. tuberculosis*. These two lineages are responsible for a majority of the tuberculosis disease burden, with Lineage 4 being the most widespread and Lineage 2 being often associated with the emergence of MDR and XDR strains (Stucki et al. 2016).

We use the gene-by-gene method, implemented as multi-locus sequence typing (MLST) to analyse the genomic variation within this species complex. MLST is scalable, readily expanding to the thousands of isolates and thousands of loci, representing specific genes or gene fragments. The database structure which forms the basis of this method ensures that it is portable and

permanent. The addition of new loci and new isolates does not require re-analysis of previously recorded data, making it readily expandable (Maiden et al. 2013).

While MLST has been used for phylogenetic characterisation of MTBC isolates in the past (Kohl et al. 2014), its potential as the foundation for an analysis of population-level signals of evolutionary pressure remains untapped.

Methods

Loci

We utilised the BIGSdb (Bacterial Isolate Genome Sequence Database) software, which uses the combination of two databases, one for *Isolates* and another for S*equence Definitions*, to analyse large datasets of bacterial genomes in a scalable way (Maiden and Jolley 2010). This software allows for an expansion of multi-locus sequence typing that can handle thousands of loci and isolates.

wgMLST: whole genome MLST

For this analysis of MTBC isolates, loci were defined based on the second revision of the H37Rv reference genome. This is the "canonical" *Mycobacterium tuberculosis* genome, and has been extensively studied since its sequencing in 1998 (Cole et al. 1998), with revisions published in 2002 (Camus et al. 2002) and 2011 (Lew et al. 2011). This annotation includes 4,090 genes, out of which 3,975 were protein-coding genes not already existing in the universal ribosomal protein database. Eight loci were added as individual paralogues of ribosomal genes. These 3,983 genes were added as new loci records in the S*equence Definitions* database for the *Mycobacterium* genus in PubMLST, each with a unique locus identifier in the MYCO000000 format. While the integers in the locus identifier are arbitrary, where possible the six-digit in the ID correlates to the Rv number in the H37Rv reference. For example, *Rv3770c* becomes MYCO003770, while *Rv3770A* and *Rv3770B* becomes MYCO013770 and MYCO023770B respectively. These loci make up the whole genome MLST scheme, or wgMLST. The first allele added to the *Sequence Definitions* database for each locus is the "seed sequence", which becomes the reference against which new alleles at that locus are compared. Here we used the alleles present in the H37Rv reference genome as the seed sequences. The H37Rv reference annotation was additionally used to characterise loci into twelve functional categories.

rMLST: ribosomal MLST

The ribosomal MLST (rMLST) scheme includes 53 genes, all of which code for protein subunits of the ribosome. This scheme is universal to all bacteria, though not all species have all 53 genes, and some species have duplications of certain genes (Maiden et al. 2012). Both of these circumstances exist in the MTBC, where BACT000021 (*rpsU*) is entirely absent, and BACT000014 (*rpsN*), BACT000018 (*rpsR*), BACT000057 (*rpmB*) and BACT000062 (*rpmG*) are duplicated. These were added as individual loci to the *Mycobacterium* 'reference sequences' database, as MYCO000717 and MYCO002056 (*rpsN1* and *rpsN2*); MYCO000055 and MYCO002055 (*rpsR1* and *rpsR2*); MYCO002057 and MYCO020634 (*rpmG1* and *rpmG2*); MYCO000105 and MYCO002058 (*rpmB1* and *rpmB2*), with the integers in the locus identifiers matching the Rv numbers of the genes. The remaining 48 loci in the rMLST scheme were combined with the wgMLST scheme to form a complete scheme of the *M. tuberculosis* genome. The rMLST scheme and its associated ribosomal sequence types (rSTs) were also the basis for the species names given to isolates.

cgMLST: core genome MLST

A core genome MLST scheme (cgMLST) was developed for the *Mycobacterium* genus, for use in the classification of isolates into species and lineages. It is composed to 749 loci. These are loci from the H37Rv reference genome that were also found to be present in 19 finished genomes of other *Mycobacterium* species, with at least 70% sequence identity and 70% alignment length. These finished reference genomes and their species are listed in Table 1, and include members of the MTBC, the *M. avium* species complex (MAC), the agent of leprosy *M. leprae* and environmental *Mycobacterium* species such as *M. abscessus*.

Table 1. Finished genomes used in *Mycobacterium* genus core scheme creation pool

Species	**Isolate name**	**GenBank ID**	**Length (in base pairs)**
Mycobacterium abscessus	ATCC 19977	NC_010397	5,067,172
Mycobacterium africanum	GM041182	FR878060	4,389,314
Mycobacterium avium	104	NC_008595	5,475,491
Mycobacterium bovis	AF2122/97	NC_002945	4,345,492
Mycobacterium canettii	CIPT 140010059	NC_015848.1	4,482,059
Mycobacterium chubuense	NBB4	NC_018022.1	5,583,723
Mycobacterium gilvum	Spyr1	CP002385.1	5,547,747
Mycobacterium indicus	MTCC 9506	NC_018612.1	5,589,007
Mycobacterium intracellulare	MOTT-64	NC_016948.1	5,501,090
Mycobacterium kansasii	ATCC 12478	NC_022663.1	6,432,277
Mycobacterium leprae	TN	NC_002677	3,268,203
Mycobacterium liflandii	128FX	NC_020133.1	6,208,955
Mycobacterium marinum	M	NC_010612	6,636,827
Mycobacterium rhodesiae	NBB3	NC_016604.1	6,415,739
Mycobacterium smegmatis	MKD8	NZ_CM001762.1	7,092,137
Mycobacterium tuberculosis	H37Rv	NC_000962	4,411,532
Mycobacterium ulcerans	Agy99	NC_008611	5,631,606
Mycobacterium vanbaalenii	PYR-1	NC_008726	6,491,865
Mycobacterium yongonense	05-1390	NC_021715.1	5,521,023

Total loci in analysis

The total number of loci included is this analysis is 3,795, as detailed in Table 1. A total possible of 4,031 loci were considered, including 3,983 loci defined in the *Mycobacterium* genus *Sequence Definitions* database from the H37Rv reference and 48 loci defined in the universal ribosomal protein database. The 147 loci that code for insertion sequences or bacteriophages, as defined in the H37Rv annotation, were excluded, as were 65 loci with designations in fewer than 15% of isolates, and 24 loci that were not complete coding sequences, that is, did not begin with a start codon and end with a stop codon with no internal stop codons.

Table 2. Loci used in this analysis and their functional categories according to the H37Rv annotation, and their abbreviations.

Functional category in H37Rv reference annotation	**Abbreviation used here**	**Loci**
Conserved hypothetical	Hypoth.	1,029
Intermediary metabolism and respiration	Int. metab.	924
Cell wall and cell processes	Cell wall	764
Lipid metabolism	Lipids	265
Virulence, detoxification and adaptation	Virulence	235
Regulatory proteins	Regulatory	198
Information pathways	Information	183
PE/PPE	PE/PPE	128
Ribosomal protein subunits	rMLST	55
Unknown	Unknown	14
Total		*3,795*

Isolates

The PubMLST Mycobacterium *Isolates* database is populated with whole genome sequences of *Mycobacterium* genus isolates obtained from the NCBI Reference Genomes database, Sequence Read Archive database (SRA) and a collection of 287 whole genomes that make up a reference collection from the *Mycobacterium* lab at Forschungszentrum Borstel (FZB).

The *Mycobacterium* genus PubMLST database contains 9,424 isolates as of August 2017, the time of writing. Out of these, 7,878 isolates belong to one of the species in the *Mycobacterium tuberculosis* species complex. Isolates for which less than 90% of the loci the *Mycobacterium* genus cgMLST scheme had allele designations were removed from this analysis, as were 2 isolates contaminated with sequences from species in other genera. The remaining 6,679 isolates were included in this analysis. This collection, as shown in Table 3, includes 6,508 *M. tuberculosis* (97.44%), 129 *M. bovis* (1.93%), 18 *M. canettii* (0.27%), 15 *M. africanum* (0.22%), 3 *M. caprae* (0.04%), 3 *M. microti* (0.04%), 2 *M. pinnipedii* (0.03%) and 1 *M. orygis* (0.01%) isolates. There are 95 finished genomes, with a single contig, and 6,584 assembled draft genomes.

Table 3. Species and genome status of isolates included in analysis.

	Genome status		
Species	**Draft**	**Finished**	**Total**
Mycobacterium africanum	14	1	15
Mycobacterium bovis	124	5	129
Mycobacterium canettii	13	5	18
Mycobacterium caprae	3	0	3
Mycobacterium microti	3	0	3
Mycobacterium orygis	1	0	1
Mycobacterium pinnipedii	2	0	2
Mycobacterium tuberculosis	6,424	84	6,508
Total	6,584	95	6,679

Genome assembly

Draft genomes were assembled using a Velvet-optimiser pipeline. Short read sequences from whole genome sequencing, with reads between 50 and 200 base pairs in length, were assembled using a pipeline that performs Velvet assemblies at all possible odd k-mer lengths between 21 (a minimum long enough to cover sections with short repeats) and the read length. The most complete assembly, measured by its total read length, a function of contig numbers and their lengths, was stored in the *isolates* database of PubMLST and used for further analyses. Assemblies varied in quality, with the number of contigs ranging from 18-1922, with a mean of 250, as noted in Table 4 and shown in Figure 1.

Table 4. The minimum, maximum and mean values for assembly statistics of the draft genomes.

	Contigs	**N50 Contig Length**	**N50 Contig Number**
Minimum	18	11,492 bp	9
Mean	250	64,579 bp	92
Maximum	1922	484,455 bp	369

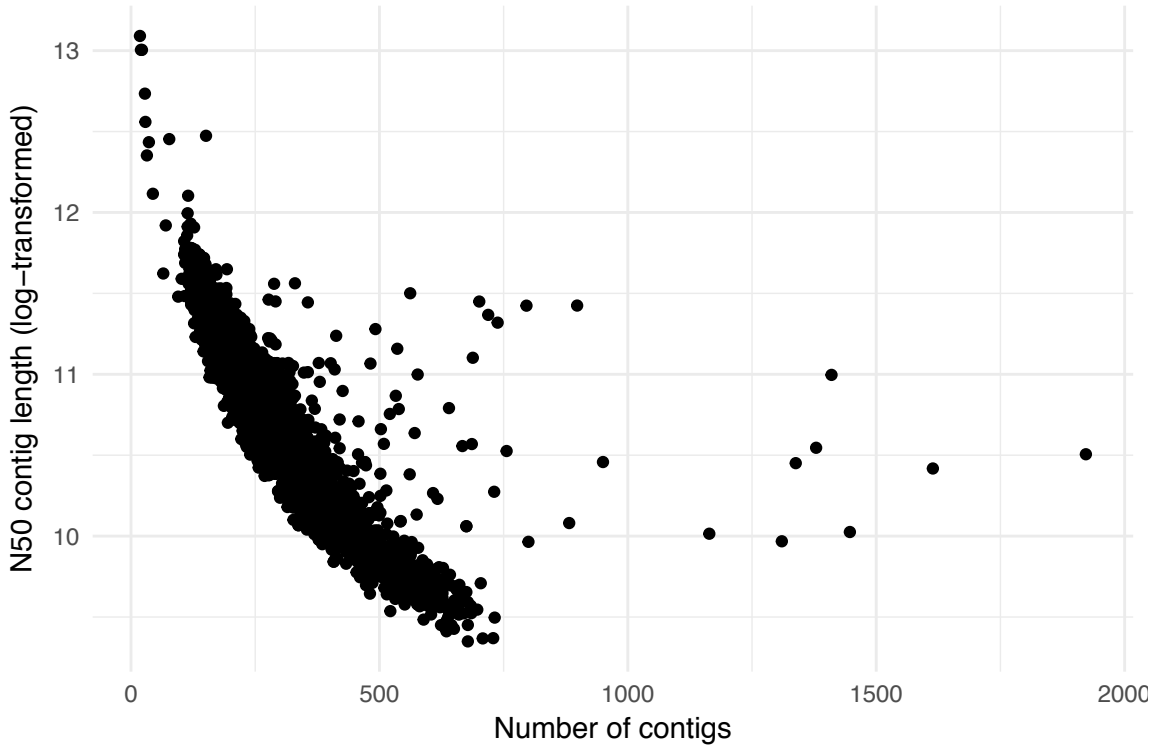


Figure 1. The relationship between contig number and N50 contig length for assembled draft genomes.

Genome Annotation

Genomes were annotated using a combination of automated and manual curation features of the BIGSdb system, as implemented on the PubMLST website. The contigs of each isolate (a single contig in the case of finished genomes, as MTBC genomes are composed of a single chromosome) are stored in the *Isolates* database. They were BLASTed against the sequences present in the '*reference sequences*' database. The criteria for a successful BLAST hit are as follows: 90% sequence identity over 90% alignment over the reference sequence, and each allele is a full coding sequence, beginning with a start codon and ending in a stop codon, with no internal stop codons. Exact matches are recorded as an "allele designation" in the *Isolates* database, while inexact matches are added as new alleles to the *Sequence Definitions* database. The iterative cycle of BLASTing against the *Sequence Definitions* database and adding sequences

to it as new alleles are found creates an annotation system that refines itself as more curation is completed.

For 21 loci with recent duplications in the genome, probes were also defined in the database to separate loci that would otherwise overlap in their successful BLAST hits. For these loci, only hits that both satisfied the BLAST conditions and occurred within 6 base pairs of the specified probe sequence were considered matches. Because both members of a paralog pair may satisfy the identity and alignment BLAST conditions, only alleles that match those requirements and additionally are in the specified genome location are added as new alleles to the '*reference sequences*' database and used for allele designations. These loci are: MYCO000277, MYCO000749, MYCO000814, MYCO000815, MYCO001037, MYCO001038, MYCO001196, MYCO001197, MYCO001198, MYCO001361, MYCO001793, MYCO002346, MYCO002347, MYCO002352, MYCO002828, MYCO003117, MYCO003118, MYCO003478, MYCO003619, MYCO003620 and MYCO003630.

Manual curation further enhances the quality of the annotation. Loci that were assigned in fewer than half of all isolates were manually investigated. In many cases this was due to major mutations in the sequences, such as gene fusions, frameshifts or deletions that removed the start or stop codon. Occasionally, an insertion or deletion created an allele whose alignment with the reference sequences was below 90%. In these situations, one allele that included that mutation was manually added to the '*reference sequences*' database. This additionally allows further alleles to be defined based on a BLAST match for that novel allele, even during automatic curation runs.

Lineage assignment

Core genome sequence types, or cgSTs, were assigned to isolates based on their core genome profiles, the combination of allele designations over the 749 loci in the cgMLST scheme. These cgST profiles were created with up to 50 loci left with no allele designations assigned, so as to be

tolerant to missing allele designations in isolates. With this tolerance, 6,281 isolates (94.04%) were assigned at least one cgST.

The classification scheme capacity of BIGSdb was used to cluster cgSTs, and therefore group the isolates assigned cgSTs. Pairwise allelic mismatches were used to cluster cgSTs, at levels of similarity determined by a maximum number of mismatches. Two such classification schemes were used to group cgSTs (and therefore isolates), with thresholds set at 120 and 75 allelic mismatches out of the 749 loci in the cgMLST scheme. A single-linkage algorithm was used to build the classification scheme groups, so that cgSTs were added to a group as long if they were below the mismatch threshold with at least one existing member of that group. This creates an effect where, given a limited population to draw isolates from, with larger numbers of isolates, fewer groups will exist, as new isolates are increasingly likely to be a match for an existing group, and groups may merge if an new isolate has mismatches below the threshold with members of two groups. Therefore these classification scheme groups are transitory, in contrast to the permanent allele identifier and sequence type profiles. Classification scheme groups were characterised as belonging to specific lineages of the MTBC if they included at least one member of the FZB reference collection as a member. Groups with a single member, or with no FZB reference collection members, were excluded.

Phylogenetic tree

Distances between isolates were calculated using their allele designations at the 749 loci of the cgMLST scheme. These distances were compiled into a dissimilarity matrix. Pairwise comparisons ignored loci where one of the two isolates is missing an allele designation. A neighbour-joining tree was created from this matrix, and labelled with each isolate lineage assignment.

Measures of diversity, selection and homoplasy

Three measures of genetic diversity and two measures of selection were calculated from the allele sequences of each of the 3,795 loci in the analysis, using the GUAVA application (detailed in the previous chapter). Two measures, one of diversity and one of selection, were allele-based, while three others were based on aligned nucleotide sequences.

Allelic diversity is an allele-based genetic diversity measure. It is analogous to SNP density, and named AllelicDiv in GUAVA. It was measured for each locus by dividing the number of unique alleles observed by the average length in base pairs of those alleles.

Allele ratio is an allele-based measure of selection. It is analogous to dN/dS ratio, and named "RatioCount" in GUAVA. It was calculated for each locus by dividing the number of unique allele sequences after translation (in amino acid format) to the number of unique allele sequences before translation (nucleotide format).

The **proportion of variable sites** in the **nucleotide** multiple sequence alignment (MSA) is an alignment-based measure of genetic diversity. It is named VSitesNuc in GUAVA. It was calculated by aligning all unique allele sequences at each locus using the FFT-NS-1 method of MAFFT (Katoh and Standley 2013), and measuring the numbers of sites in the alignment which had any variation and the alignment length. These are used to calculate the proportion of the alignment that is made of variable sites.

The **proportion of variable sites** in the **amino acid** multiple sequence alignment (MSA) is also an alignment-based measure of genetic diversity. It is named VSitesAA in GUAVA. It is similar to the above, but calculated from an amino acid alignment instead, so variation is limited to mutations that cause non-synonymous changes. Like the allele-based measure, it is calculated

from an alignment of all unique allele sequences, in this case in amino acid format, and measuring the proportion of variables sites in the alignment from this.

The **ratio of proportions** is an alignment-based measure of selection. It is named RatioVS in GUAVA. Like the allele ratio, it is analogous to the dN/dS ratio, in that it also measures the ratio of non-synonymous to synonymous mutations. It is simply the ratio of the proportion of variable sites in the nucleotide MSA to proportion in the amino acid MSA.

To calculate p-values, all variables, with the exception of the allele ratio, were log-transformed to obtain a normal distribution, and then transformed into a measure of their deviation from their collective mean value to obtain their z-score. These were used to calculate the probability of each value occurring in a population with the mean and standard deviation of the samples from that variable. These probabilities were corrected for multiple testing using the Holm method. The p-values for these measures therefore represent the degree to which the locus in question is "aberrant" in contrast to other loci in the genome, given the patterns of genetic variation observed in the population.

To calculate the correlation between the essentiality probabilities measured by Griffin et al through mutagenesis and deep-sequencing (Griffin et al. 2011) to the allele ratio, a linear regression model was used.

Two lineages, named Lineage 2 and Lineage 4, were investigated for their largest differences in the allele-based measures (allelic diversity and allele ratio) to other lineages of the MTBC. Each measure was calculated separately for only Lineage 2 isolates, and for only isolates in lineages other than Lineage 2. The ratio between these values obtained was calculated, and these ratios were used as the basis for p-value calculations as before. This was repeated using Lineage 4 and all isolates assigned to lineages other than Lineage 4.

Loci with high levels of homoplasy were identified by shared allele presence. At each locus, the lists of unique alleles present within Lineage 2 and Lineage 4 were contrasted. The count of unique alleles over the entire locus as well as the count of alleles that appeared within both lineages was measured. Loci with more than 5 alleles appearing in both lineages and where "shared" alleles represented 15% or more of the total number of alleles were determined to display high levels of homoplasy.

Results

Curation of isolates and loci

With a dataset of 6,679 isolates and 3,860 loci, a potential maximum of 25,780,940 allele designations can be made, and 24,403,680 (94.66%) were assigned. Most loci had designations in the majority of isolates, as can be seen in Figure 2. The remaining unassigned designations represent a mixture of alleles with mutations that cause them to no longer be complete coding sequences, such as internal stop codons, frameshifts or deletions, alleles left incomplete due not being within a contig after genome assembly, and possibly some alleles missed due to incomplete curation. Loci that had designations in fewer than 15% of isolates were removed from the analysis as statistical tests are skewed by unequal sampling and generally have lower power with fewer samples, so that 3,795 loci are investigated in this study. Out of all allele designations made, 19,245,191 (74.65%) are for the allele with identifier "1", which denotes an exact match to the seed sequence from the H37Rv reference genome.

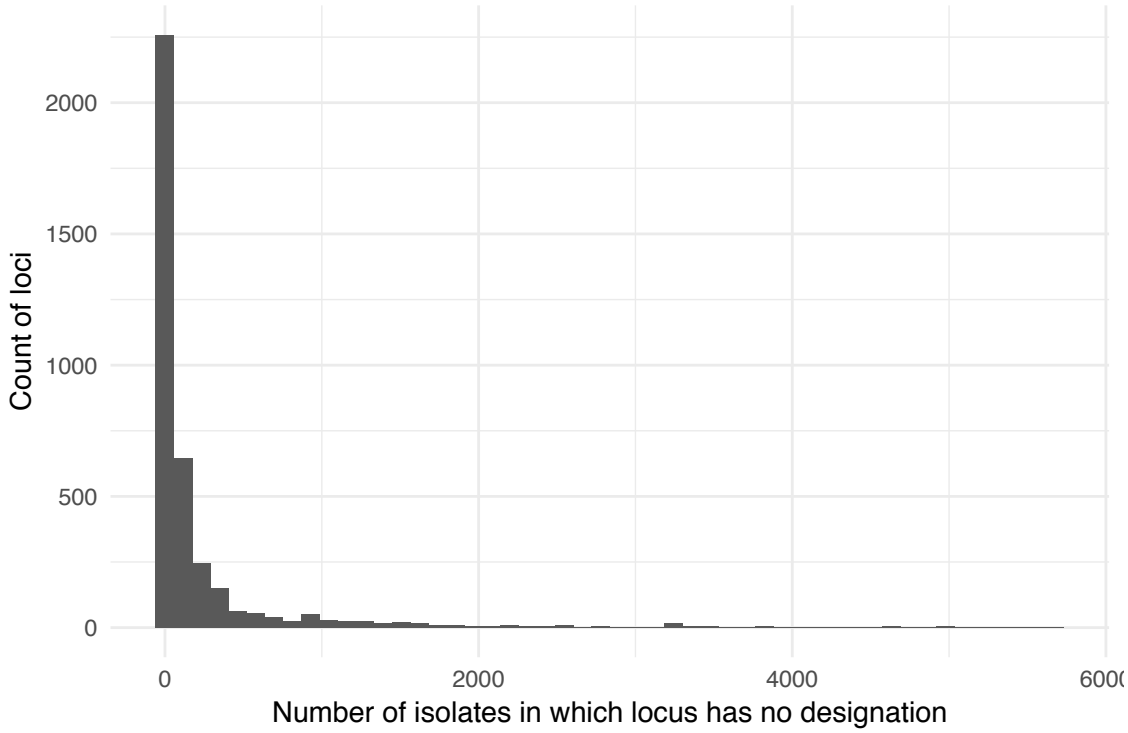


Figure 2. Distribution of loci according to the number of isolates in which they had no allele designations.

Lineage assignment

Out of the 6,679 isolates included in this analysis, 6,298 (94.3%) isolates had cgMLST profiles complete enough to be assigned a cgST. There were 5,625 cgSTs assigned to isolates, with the most frequent being assigned to 190 isolates. A large proportion of cgSTs, 4,891 or 86.95%, were assigned to a single isolate.

Classification by a 120 maximum allelic mismatch threshold created 27 groups. Ten of these included at least one member from the FZB reference collection and could therefore be assigned to known lineages of the MTBC. These included one group representing the "modern" lineages of *M. tuberculosis*, one for Lineage 1, one for Lineage 5, two for Lineage 6 and four for various

animal-adapted MTBC species. Seventeen groups had only a single member. The largest group, with 5,481 isolates (87.03%) included members of multiple MTBC lineages, and so was further characterised by a classification scheme with a threshold of 75 allelic mismatches. In this classification scheme, the 5,481 isolates were clustered into 16 groups. Three groups corresponded to sub-lineage of Lineage 4, two groups to sub-lineages of Lineage 2 and two groups to sub-lineages of Lineage 3. Six groups of which had a single member, and three groups had no FZB reference collection isolates, as detailed in Table 6.

Classification groups could therefore be used to assign lineages to 6,268 isolates. A further 12 isolates could be characterised despite having no cgST assigned to them as they were members of the FZB reference collection. Therefore 6,280 (94.03%) isolates could be assigned to a lineage, as described in Table 7.

Table 5. Groups in the cgMLST classification scheme with a threshold of 120 allele mismatches.

Group ID	Description	Isolates
1	"Modern" lineages (2, 3 and 4)	5,481
3	*M. bovis*	135
5	Lineage 5: West African 1a or 1b	57
6	Lineage 6: West African 2	56
9	Lineage 1: EAI	528
15	Lineage 6: animals	9
23	*M. caprae*	2
24	*M. canettii*	2
25	*M. canettii*	6
26	Animal strains (vole, seal, etc.)	5
—	In groups with only one member	17
	Total	*6,298*

Table 6. Groups in the classification scheme with a threshold of 75 allele mismatches, found within Group 1 of the classification scheme with a threshold of 120 allele mismatches.

Group ID	Description	Isolates
1	Lineage 4: Ghana, Haarlem, X-type	588
2	Lineage 4: Cameroon, LAM, Uganda, New-1, S-type	2,438
10	Lineage 2: Beijing	1,375
12	Lineage 4: TUR, Ural	171
13	Lineage 3: Delhi-CAS	883
33	Lineage 2	11
56	Lineage 3: Delhi-CAS	2
—	In groups with no reference members (3 groups)	7
—	In groups with only one member	6
	Total	*5,481*

Table 7. Lineage assignments for all isolates in which this was possible.

Lineage	**Isolates**
Animal strains	151
Lineage 1	531
Lineage 2	1,387
Lineage 3	887
Lineage 4	3,200
Lineage 5	58
Lineage 6	66
Total	*6,280*

Phylogenetic structure

A neighbour-joining phylogenetic tree was constructed for the MTBC, excluding *M. canettii*, based on allele mismatches in the cgMLST scheme, and labelled with lineage assignments, as shown in Figure 3. The pairwise allele mismatch distances from the matrix show an uneven distribution with several distinct peaks, as shown in Figure 4. Distances between members of the same lineage varied by lineage, as shown in Figure 5 and Table 8. Lineage 6 and *M. canettii* had the most genetically diverse set of members, with a maximum mismatch distances of 294.4 and 294.3 alleles respectively. The three "modern" lineages of M. tuberculosis showed the lowest diversity, with maximum pairwise mismatches of 143.3 (Lineage 2), 149.4 (Lineage 3) and 162.8 (Lineage 4).

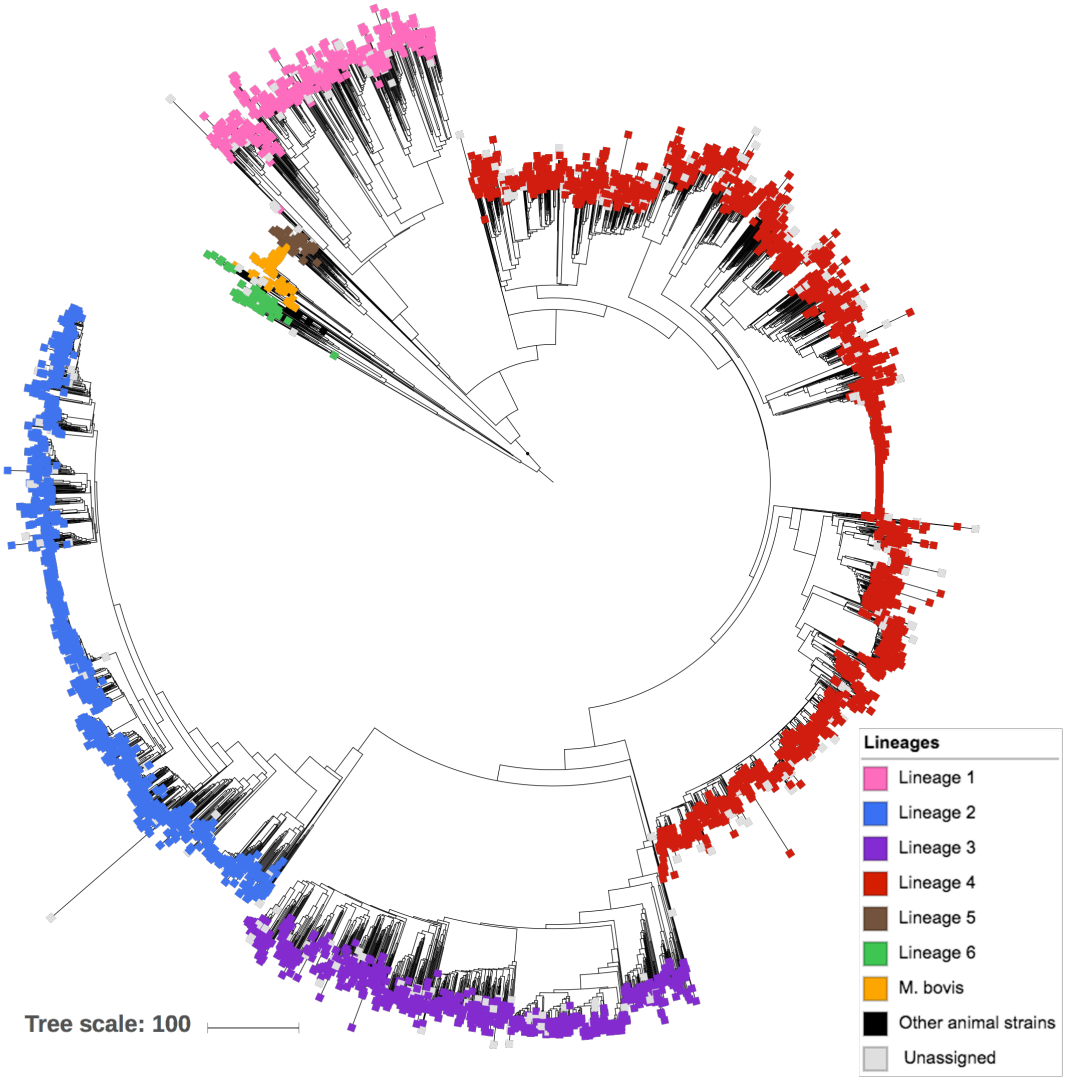


Figure 3. A Neighbour-Joining phylogenetic tree of 6,661 isolates in the MTBC. *M. canettii* isolates were excluded due to their large distance to other lineages. Distances between isolates were calculated as a dissimilarity matrix based on allele matches at each of the 749 loci of the cgMLST scheme, with pairs where one or both isolates had no designations being excluded from that pairwise calculation. Tips are coloured based on their assigned lineages, with grey marking isolates that were not assigned to any lineages.

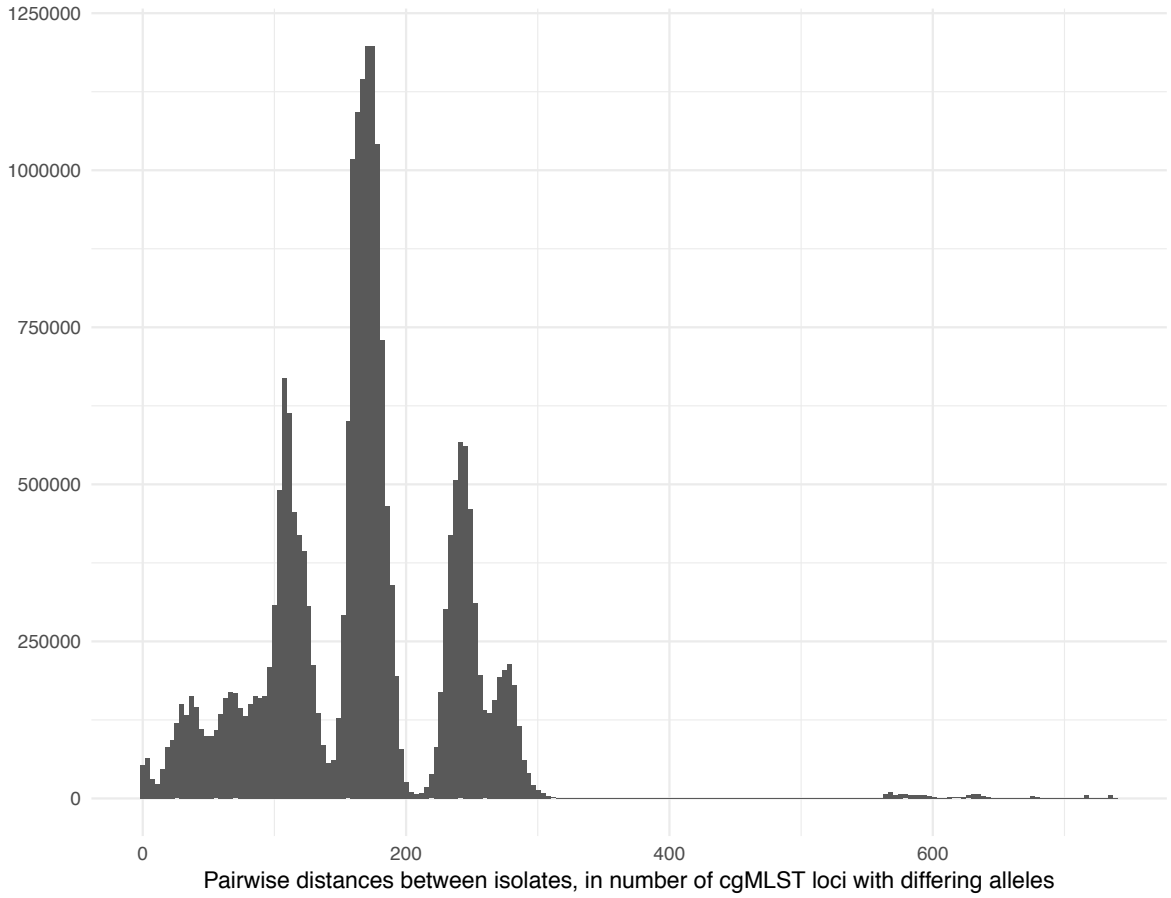


Figure 4. Pairwise distances between 6,679 MTBC isolates, across 749 loci in the *Mycobacterium* genus cgMLST scheme.

The pairwise allele mismatch distances from the matrix show an uneven distribution with several distinct peaks, caused by the phylogenetic structure of the MTBC population. Pairwise comparisons that include *M. canettii* isolates form the outliers seen at the right-hand side of the graph, with very large genetic distances compared to any other pairs within the MTBC.

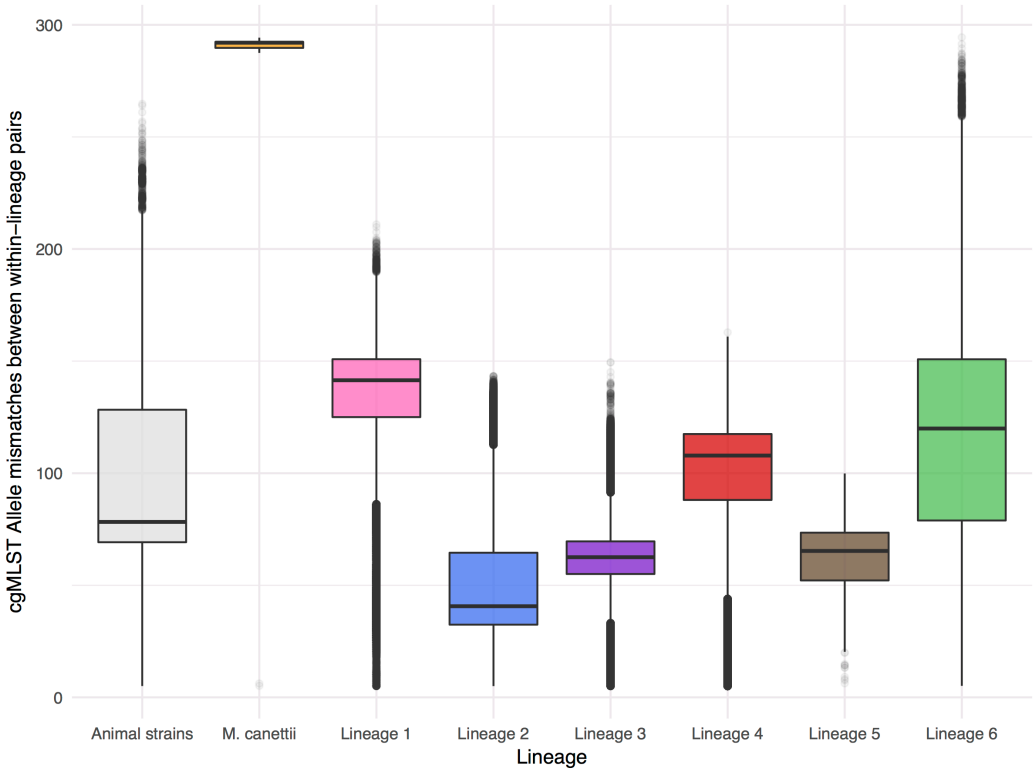


Figure 5. Pairwise distances within the major lineages of the MTBC. Pairwise comparisons of five or fewer allele mismatches were removed.

Table 8. Pairwise distances within the major lineages of the MTBC and in total for the species complex.

	Mean pairwise mismatch distance	**Maximum pairwise mismatch distance**
All isolates	90.9	294.4
Animal strains	99.4	264.7
M. canettii	251.0	294.3
Lineage 1	130.5	211.0
Lineage 2	51.2	143.3
Lineage 3	62.3	149.4
Lineage 4	99.4	162.8
Lineage 5	62.6	99.9
Lineage 6	139.9	294.4

Signals of diversity and selection

Allelic diversity (measured as the count of unique alleles divided by the average allele length in base pairs to control for varying locus lengths) was calculated for all 3,795 loci as a measure of genetic diversity. Allelic diversity had a mean value of 0.0614 alleles per nucleotide, with a minimum of 0.0090, a maximum of 0.3822 and a standard deviation 0.0187. Fourteen loci had allelic diversity values significantly differing from the remainder of the population, as detailed in Table 9 and shown in Figure 6.

Six loci had values of allelic diversity significantly below the average. Four of these are unknown function, namely *Rv0197* ($p = 0.03914$), *Rv2084* ($p = 0.00926$), *Rv2970A* ($p = 0.00346$) and *Rv3453* ($p < 0.0001$). Another, *aceAa* ($p = 0.00646$) may have a low level of observed diversity values because one of its possible mutations is a fusion with *aceAb*, the locus downstream. This has been observed *M. tuberculosis* reference genomes other than H37Rv, such as CSU93 . Lastly, *pks15* ($p = 0.01262$) also displays a low level of genetic diversity. Its product is a polyketide synthase involved in the production of phenolic glycolipid (PGL-tb), a virulence marker for *M. tuberculosis*. This locus typically contains a deletion causing a frameshift, with the exception of Lineage 2, whose members can produce PGL-tb as *pks15* is intact (Sinsimer et al. 2008). This restriction to a single lineage is likely to be the source of its observed low level of genetic diversity.

Eight loci had significantly elevated allelic diversity. A majority of these had functions related to virulence. Two of these are known to be involved in drug resistance, namely *gid* ($p < 0.0001$) that may confer resistance to streptomycin and *pncA* ($p < 0.0001$) that may do the same for pyrazinamide. Three highly diverse loci had functions related to the ESAT-6 like proteins, out of which the two are *esxO* and *esxN* ($p < 0.0001$), both of which have putative ESAT-6-like proteins as products, and *whiB6* ($p < 0.0001$), which is known to be essential for and able to modulate the

expression of *esxA* and *esxB* (Chen et al. 2016). Lastly, *vapC25* ($p = 0.0002$) is a possible toxin, and *PPE60* ($p < 0.0001$) and *Rv0010c* ($p = 0.0021$) have unknown functions.

The presence of loci related to cell wall proteins in both extremes of this measure may be due to the role of the cell wall as the boundary between the bacterial cell and the host. This makes the cell wall an essential structure, and some of the most conserved proteins are present there due to that crucial role. However, interaction with the host's immune system can also occur at the cell wall, and those proteins can therefore be under strong diversifying selection to avoid recognition by the immune system. For this reason, the cell wall functional category can include loci that are both notably conserved and notably diverse, and this effect is seen repeatedly through the many measures of diversity explored here.

Table 9. Loci with allelic diversity values significantly different from genome average.

Locus name	Functional Category	Unique alleles	Average length, in basepairs	Allelic diversity	p-value
Rv3453	Cell wall	3	333	0.00901	2.93E-10
Rv2970A	Hypoth.	3	171	0.01754	0.00346
aceAa	Int. metab.	20	1104	0.01812	0.00646
Rv2084	Unknown	21	1137	0.01847	0.00926
pks15	Lipids	28	1491	0.01878	0.01262
Rv0197	Int. metab.	45	2250	0.02	0.03914
gid	Cell wall	258	675	0.3822	5.29E-10
pncA	Int. metab.	190	560.9	0.3387	1.60E-08
esxO	Cell wall	96	285	0.3368	1.87E-08
esxN	Cell wall	81	285	0.2842	1.53E-06
PPE60	PE/PPE	334	1180	0.2831	1.69E-06
whiB6	Regulatory	86	351	0.245	5.06E-05
vapC25	Virulence	100	429	0.2331	0.0002
Rv0010c	Cell wall	86	417.4	0.206	0.0021

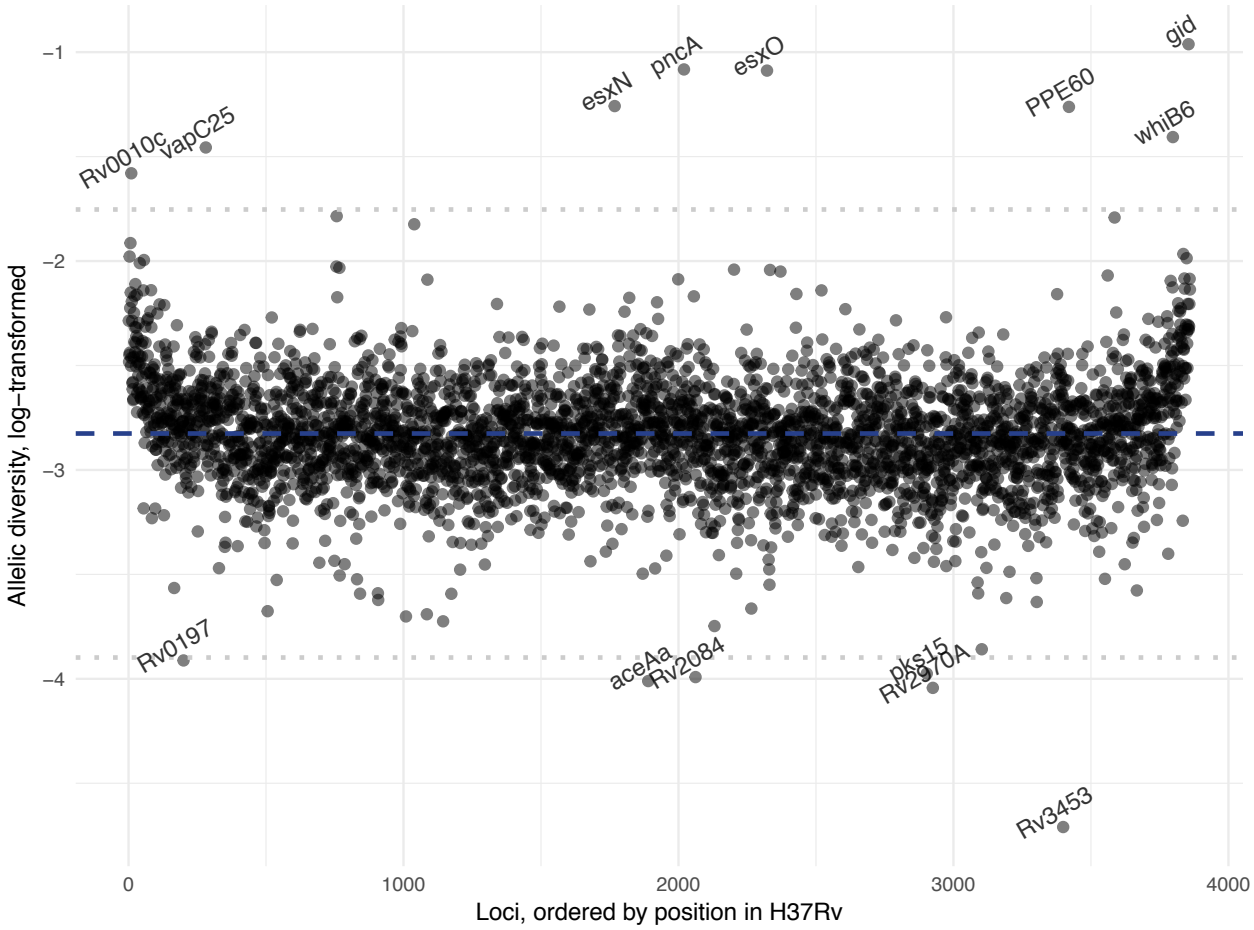


Figure 6. Allelic diversity values, log-transformed for 3,795 loci in the MTBC.

A measure of the rate of synonymous to non-synonymous mutations was acquired by taking the ratio of unique sequences after and before translation. This gives the allele ratio value, which here had a mean of 0.6443 across all loci, with a standard deviation of 0.0978 and a range of 0.2222 to 1.000, as shown in Figure 7. Two loci had the maximum value, of 1, where every allele was unique even after translation. These were *vapB6* and *Rv2970A*, two of the smallest loci, at 156 and 171 base pairs in length. Only one locus had a value significantly lower than from the genome average, namely *folE* ($p = 0.030$), as shown in Table 10. Its product, GTP cyclohydrolase I, is essential to the metabolism of the cell as it is part of the folate biosynthesis pathway.

Table 10. Loci with allele ratio values significantly different from genome average.

Locus name	**Functional Category**	**Unique alleles pre-translation**	**Unique alleles post-translation**	**Allele ratio**	**p-value**
folE	Int. metab.	18	4	0.2222	0.029923159

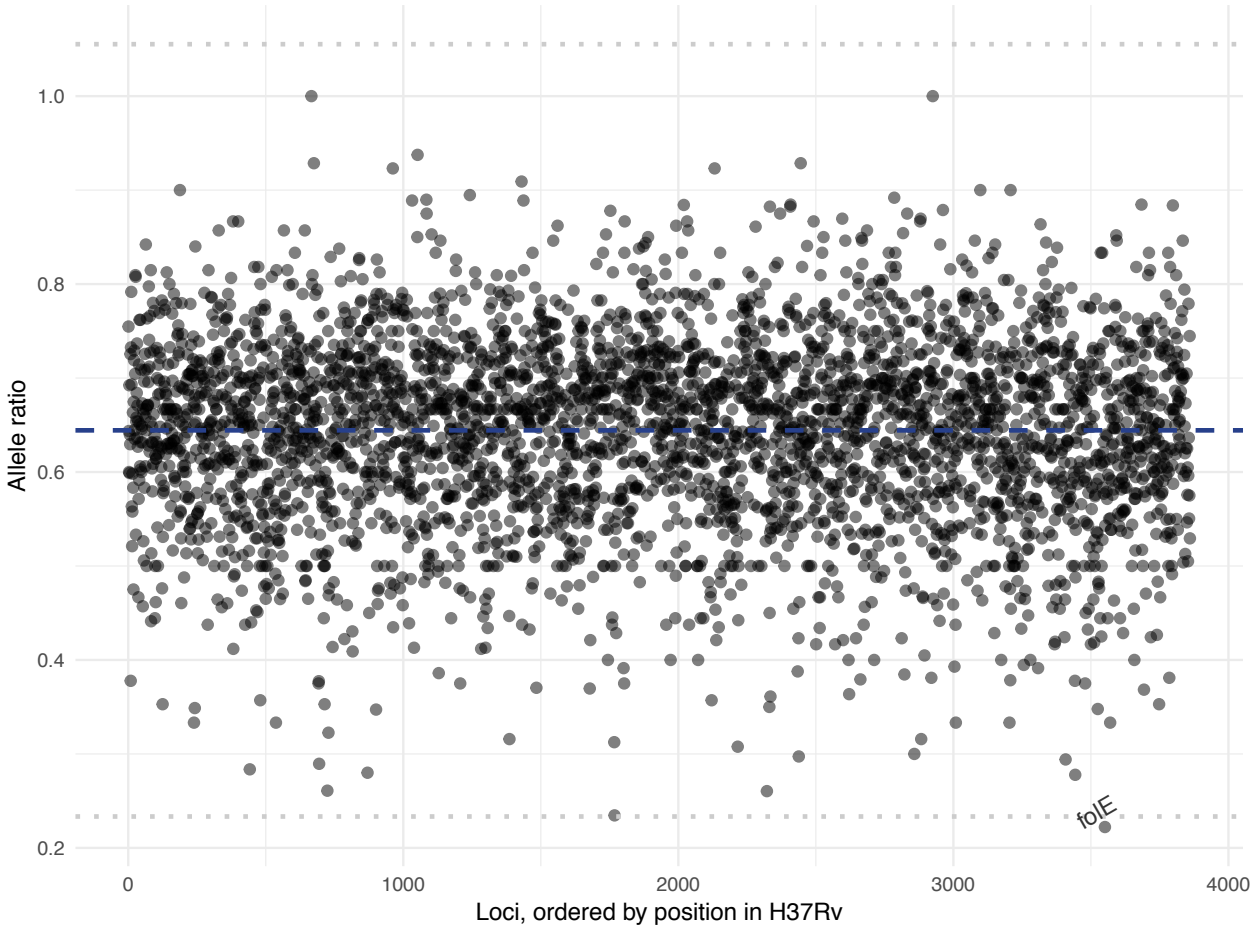


Figure 7. Allele ratio values for 3,795 loci in the MTBC.

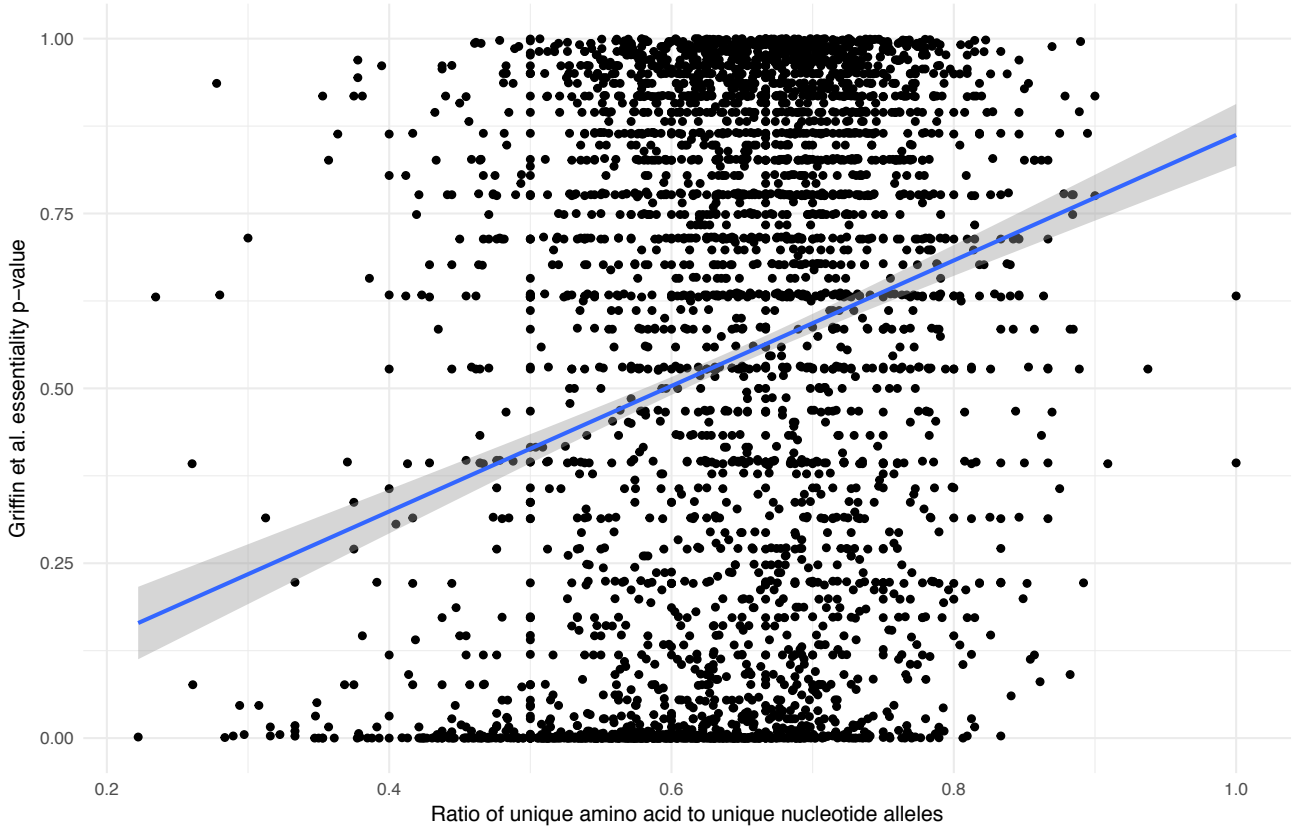


Figure 8. Correlation between Griffin et al (2016) essentiality p-value and the allele ratio.

Griffin et al. experimentally calculated the likelihoods that genes are essential for growth and replication of *M. tuberculosis* (Griffin et al. 2011). These values correlated to the proportion of amino acid sequence alignment sites with variation ($p < 2.2e\text{-}16$), with an R^2 value of 0.2404 (95% CI: 0.2101-0.2703).

The proportion of sites in a nucleotide multiple sequence alignment for all alleles at each locus had a mean value of 0.08258 across all loci, with a standard deviation of 0.03272 and a range of 0.006006 to 0.4315, as shown in Figure 9. Only five loci had proportions significantly below the genome average, all of which also have allelic diversity values significantly below average: *aceAa* ($p = 0.01707$), *pks15* ($p = 0.03391$), *Rv2084* ($p = 0.01137$), *Rv2970A* and *Rv3453* ($p < 0.0001$), as shown in Table 11. Eight loci had significantly higher proportions than average, with over 32% of the total alignment length being composed of sites with variation, indicating insertions, deletions or inversions that affect large areas of the locus. Half of these belong to the PE/PPE family of proteins unique to the *Mycobacterium* genus: *PE_PGRS13* ($p = 0.03203$), *PE_PGRS21* ($p = 0.01868$), *PE_PGRS26* ($p = 0.00938$) and *PPE16* ($p = 0.04603$). The second most populous category was cell wall and cell processes, with *rpfA* ($p = 0.00085$), one of five resuscitation-promoting factors, and two loci with unknown functions *Rv1435c* ($p = 0.00341$) and *Rv0010c* ($p = 0.01363$). A single locus related to lipid synthesis was present: *accE5* ($p = 0.00132$), which has been implicated in production of cell wall structures that resist the defense mechanisms of host organisms (Bazet Lyonnet et al. 2014).

Table 11. Loci with proportions of variable sites in a nucleotide multiple sequence alignment that differ significantly from genome average.

Locus name	Functional category	Average length, in base pairs	Length of MSA	Proportion of variable sites in MSA	p-value
Rv3453	Cell wall	333	333	0.006006	8.92E-11
Rv2970A	Hypoth.	171	171	0.0117	4.66E-05
Rv2084	Unknown	1137	1137	0.01671	0.01137
aceAa	Int. metab.	1104	1104	0.01721	0.01707
pks15	Lipids	1491	1491	0.01811	0.03391
rpfA	Cell wall	1219	1569	0.4315	0.00085
accE5	Lipids	497.6	699	0.4192	0.00132
Rv1435c	Cell wall	626.8	801	0.3933	0.00341
PE_PGRS26	PE/PPE	1455	1697	0.3665	0.00938
Rv0010c	Cell wall	417.4	426	0.3568	0.01363
PE_PGRS21	PE/PPE	2307	2567	0.3487	0.01868
PE_PGRS13	PE/PPE	2238	2403	0.335	0.03203
PPE16	PE/PPE	1850	2019	0.3259	0.04603

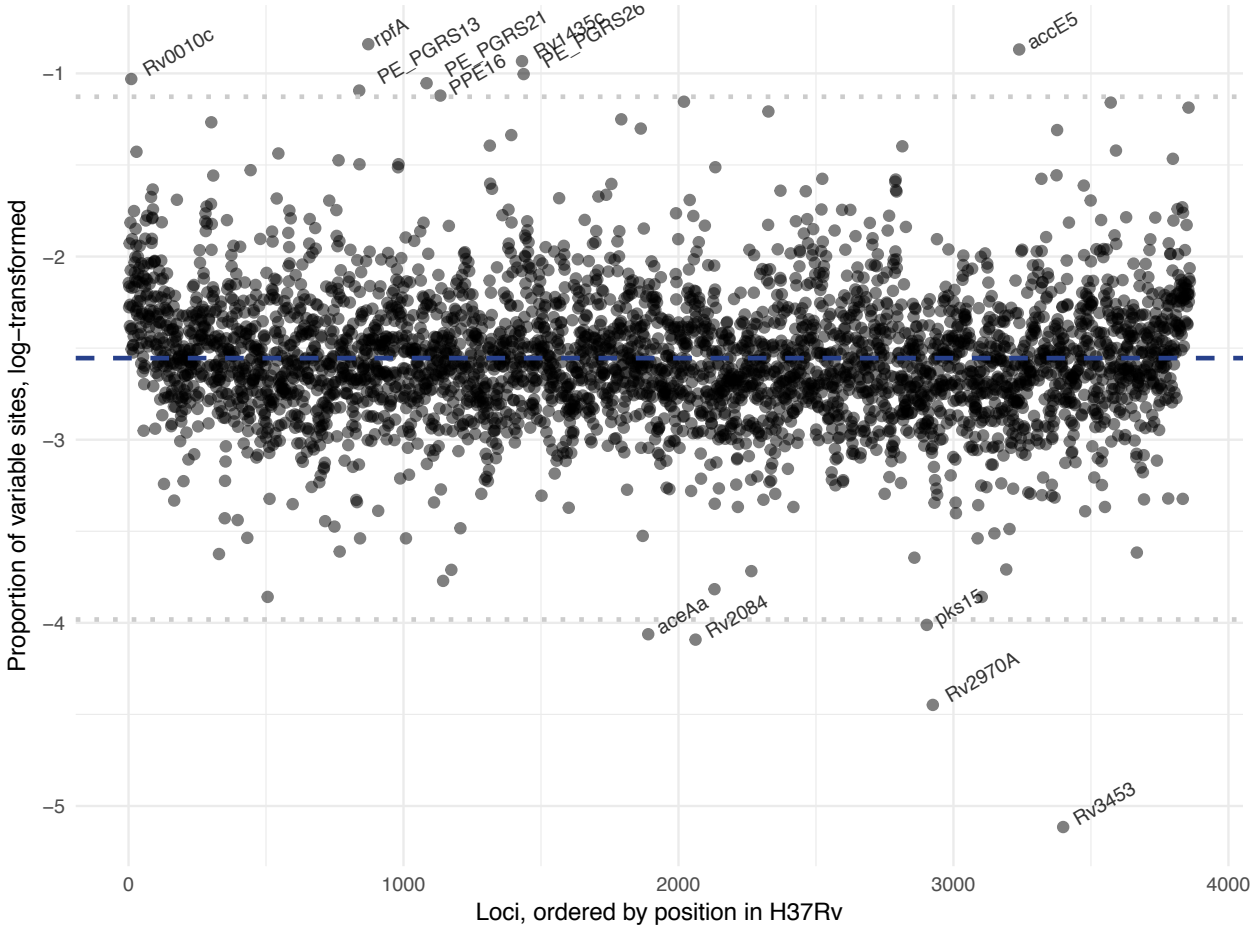


Figure 9. Proportions of variable sites in a nucleotide multiple sequence alignment, log-transformed for 3,795 loci in the MTBC.

Multiple sequence alignments of translated sequences revealed the proportion of amino acid sites with variation. These proportions which had a mean of 0.1286 across all loci, with a standard deviation of 0.05431 and a range of 0.0091-0.7347. Three loci had a significantly lower than average proportion, including *folE* ($p = 0.00016$), the locus with the lowest allele ratio, and two loci with unknown function, *Rv3453* ($p < 0.0001$) and *Rv2901c* ($p = 0.00727$). All four loci with significantly high proportions are from the cell wall and cell processes functional category, but only one, *rpfA* ($p = 0.02573$) has a known function, as a resuscitation promotion factor.

Table 12. Loci with proportions of variable sites in an amino acid multiple sequence alignment that differ significantly from genome average.

Locus name	Functional category	Average length, in base pairs	Length of MSA	Proportion of variable sites in MSA	p-value
Rv3453	Cell wall	333	110	0.0091	6.68E-08
folE	Int. metab.	609	202	0.0149	0.00016
Rv2901c	Hypoth.	306	101	0.0198	0.00727
Rv0010c	Cell wall	417.4	147	0.7347	0.00551
Rv0461	Cell wall	525	208	0.7260	0.00640
rpfA	Cell wall	1219	680	0.6471	0.02573
Rv1233c	Cell wall	596.9	228	0.6228	0.04012

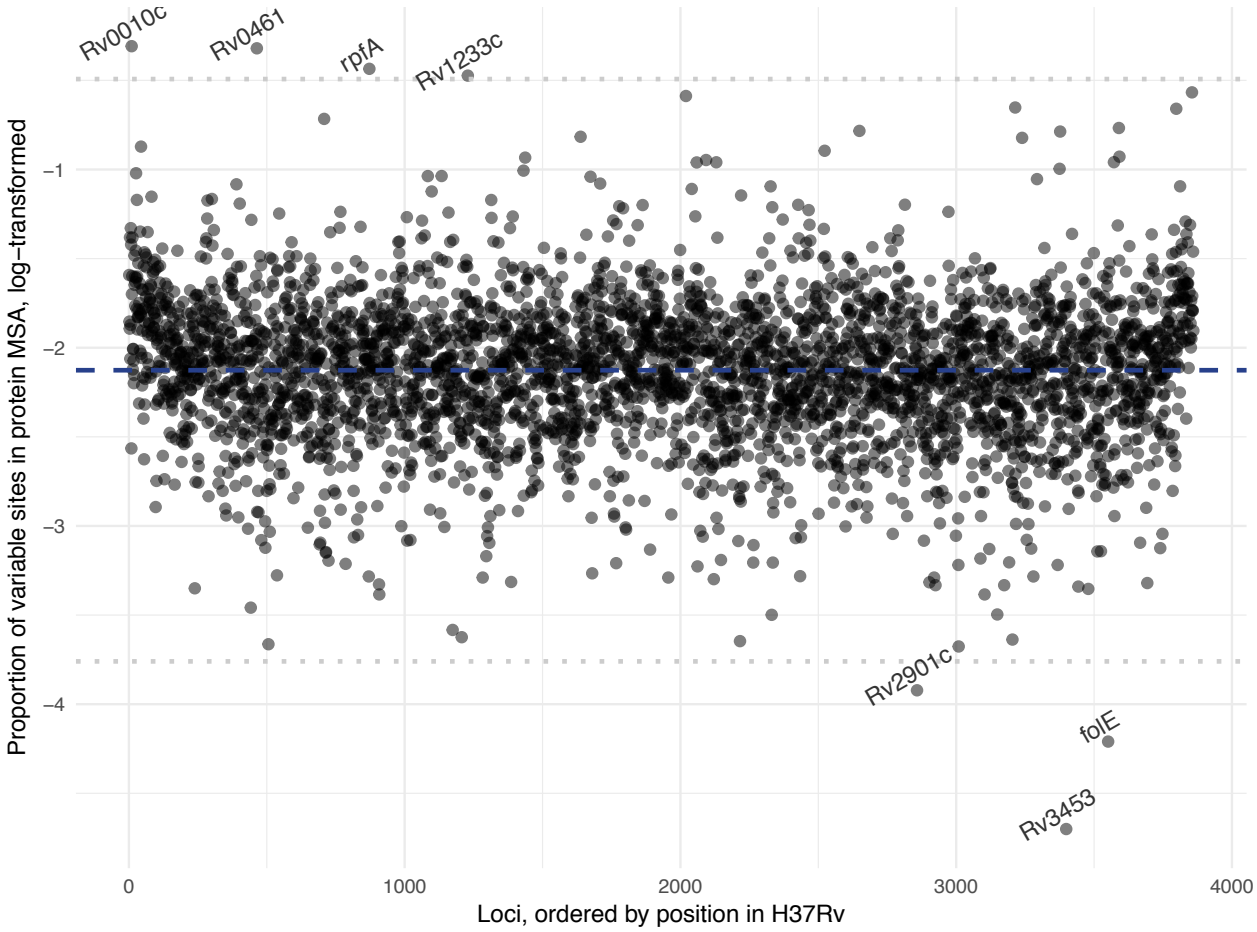


Figure 10. Proportions of variable sites in an amino acid multiple sequence alignment, log-transformed for 3,795 loci in the MTBC.

The ratio of the proportions of variable sites in amino acid and nucleotide formats was measured, giving the proportion ratio value. Its mean value was 1.5770, with a standard deviation of 0.4005, and a range of 0.4307-7.1750. Six loci had low proportion ratios, indicating a low proportion of variable sites in their amino acid multiple sequence alignment, and therefore a low ratio of synonymous to non-synonymous mutations. These include *folE* ($p < 0.0001$), which was also noted for its low allele ratio, and a second locus in the intermediary metabolism category, namely *clpX* ($p = 0.00434$), a Clp-ATPase that is essential for growth and targeted by acyldepsipeptide antibiotics. Two virulence-associated loci, *groEL2* ($p = 0.01911$) and *yrbE4A* ($p = 0.04985$) also had significantly low ratios. The product of *groEL2* is the abundant Cpn60.2 chaperonin, which blocks macrophage apoptosis and is essential for growth in the host. While *yrbEA4A* has no certain function, it is one of eight genes in an operon that is expressed in response to hypoxia and cholesterol, conditions that may be found during infection (Rathor et al. 2016). Two cell wall proteins, *secE2* ($p = 0.00105$) and *Rv3355c* ($p = 0.01696$) likewise had low ratios, but no known function. They can be stipulated from this data to have an essential cell function however, as they are strongly conserved by purifying selection. Their products likely form a structural part of the cell wall, and are especially unlikely to interact with the host immune system, as that would tend to increase their diversifying selection pressure.

Fourteen loci had allele ratios significantly above average, with four or more times higher proportions of variable sites in amino acid than nucleotide alignments, indicating a high density of non-synonymous mutations, as shown in Table 13. These included six loci without known functions, *Rv3256c* ($p < 0.0001$), *Rv0461* ($p < 0.0001$), *Rv3337* ($p < 0.0001$), *Rv1233c* ($p < 0.0001$), *Rv2248* ($p = 0.00028$) and *Rv2081c* ($p = 0.01589$). One locus classified as hypothetical in the reference genome annotation has since been characterised, as *Rv00698* ($p = 0.000968$) is now lldD1, with a function in lactate oxidisation which may be critical for metabolism within host macrophages . Two intermediary metabolism loci had high ratios, namely *argD*, involved in

arginine biosynthesis and serB1, a phosphoserine phosphatase, and one cell wall protein that is a putative arsenic transporter, *arsB1* (p = 0.0122). From the PE/PPE family of genes, two loci had significant low ratios, namely *PE_PGRS60* (p = 0.005816), which binds to fibronectin to aid host cell invasion (Meena and Meena 2016), and *PPE37* (p = 0.013229), which may interfere with the cytokine response of the host . Lastly, two loci involve in lipid metabolism showed high rates of synonymous mutations: *Rv0045c* is a hormone sensitive esterase that may be involved in dormancy (Sherlin and Anishetty 2015) and *fadE7* which is involved in cholesterol metabolism.

Table 13. Loci with a ratio of proportions that differ significantly from genome average.

Locus name	Functional category	Proportion of variable sites in amino acid MSA	Proportion of variable sites in nucleotide MSA	Ratio of proportions	p-value
folE	Int. metab.	0.03448	0.01485	0.4307	4.61E-05
secE2	Cell wall	0.1435	0.07042	0.4907	0.00105
clpX	Int. metab.	0.07182	0.03756	0.523	0.00434
Rv3355c	Cell wall	0.1293	0.07216	0.5581	0.01696
groEL2	Virulence	0.05607	0.03148	0.5614	0.01911
yrbE4A	Virulence	0.06013	0.03543	0.5892	0.04985
Rv3256c	Hypoth.	0.07259	0.5208	7.175	2.53E-08
Rv0461	Cell wall	0.1098	0.726	6.612	2.82E-07
argD	Int. metab.	0.06894	0.442	6.411	6.80E-07
Rv3337	Hypoth.	0.0567	0.3485	6.146	2.20E-06
Rv1233c	Cell wall	0.1015	0.6228	6.136	2.30E-06
Rv2248	Hypoth.	0.0625	0.318	5.088	0.000278
serB1	Int. metab.	0.0463	0.2242	4.842	0.000885
lldD1	Hypoth.	0.1013	0.4886	4.823	0.000968
Rv0045c	Lipids	0.09211	0.4182	4.54	0.003725
PE_PGRS60	PE/PPE	0.1044	0.4643	4.447	0.005816
fadE7	Lipids	0.07065	0.3038	4.3	0.011790
arsB1	Cell wall	0.1064	0.4568	4.293	0.012191
PPE37	PE/PPE	0.09072	0.3879	4.276	0.013229
Rv2081c	Cell wall3	0.09031	0.3827	4.238	0.015886

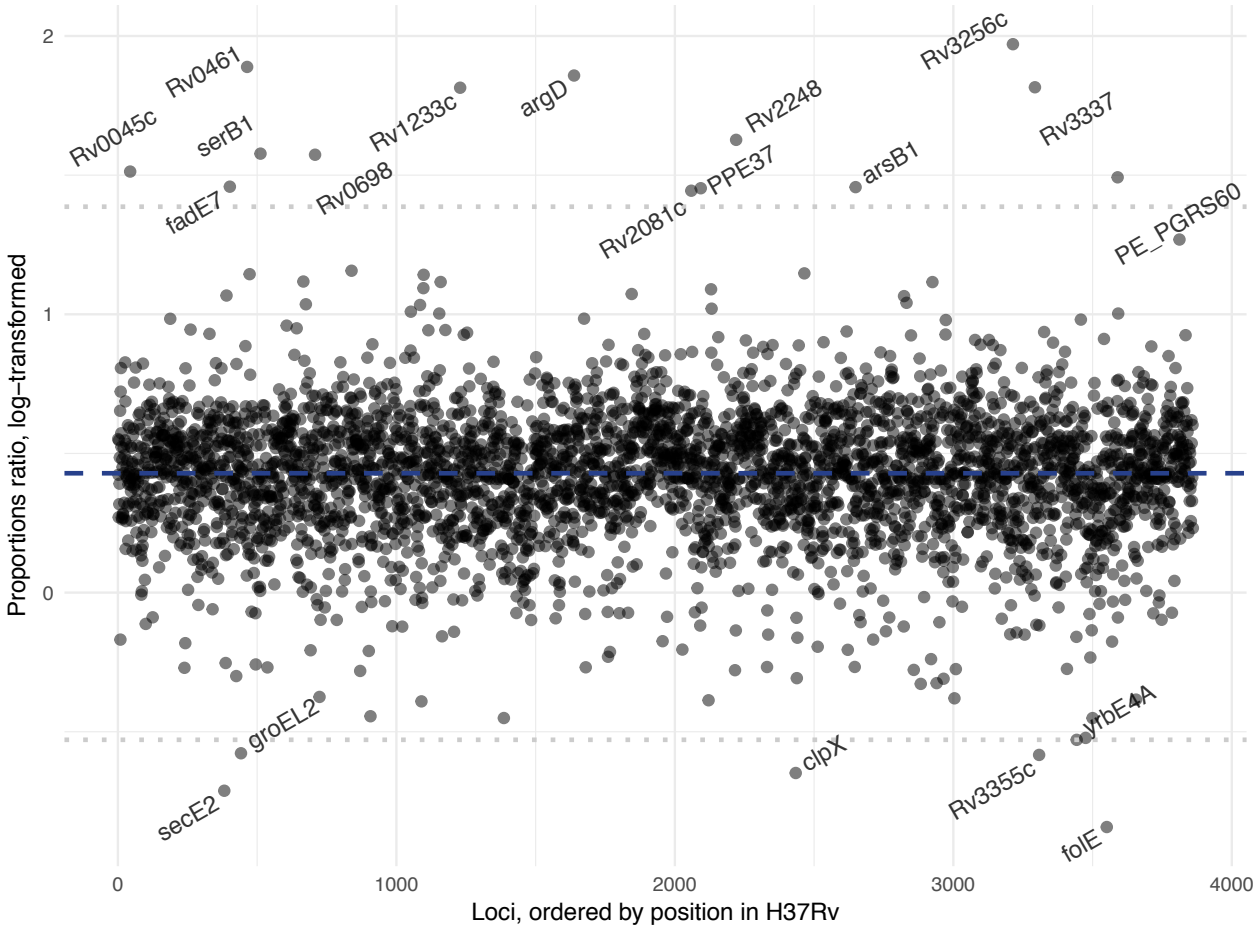


Figure 11. Ratio of proportions of variable sites in protein to nucleotide multiple sequence alignments, log-transformed for 3,795 loci in the MTBC.

Two measures, allelic diversity and allele ratio, were calculated for isolates within Lineage 2 and isolates in all other Lineages, and the differences between them was assessed to find the loci most different between those lineages and their contexts.

Six loci were found to differ significantly in their genetic diversity between Lineage 2 and other lineages, as shown in Table 14. Five of these had higher diversity outside of Lineage 2, including three loci with hypothetical products, namely *Rv0738* ($p = 0.01717$), *Rv3361c* ($p = 0.01726$) and *Rv2147c* ($p = 0.02357$). One locus, *ilvC* ($p = 0.01722$), is involved in internal metabolism, specifically valine and isoleucine biosynthesis, while another, *Rv0488* ($p = 0.03253$) has a probably role in lysine transport across the cell wall. Only one locus, *fdxC* ($p = 0.00601$) had higher diversity within Lineage 2, with six alleles within Lineage 2 and no variation (only one allele) in all others. It is a ferrodoxin, an iron-sulphur protein that transports electrons.

Table 14. Loci significantly different values of allelic diversity between Lineage 2 and other lineages.

Locus name	Functional category	Average length in base pairs	Unique alleles outside Lineage 2	Unique alleles within Lineage 2	p-value
Rv0738	Hypoth.	549	34	1	0.01717
ilvC	Int. metab.	1002	34	1	0.01722
Rv3361c	Hypoth.	552	34	1	0.01726
Rv2147c	Hypoth.	726	33	1	0.02357
Rv0488	Cell wall	606	32	1	0.03253
fdxC	Int. metab.	327	4	6	0.00601

Eight loci showed higher allele ratio values within Lineage 2 than outside it, indicating a high ratio of non-synonymous mutations within Lineage 2 especially, as shown in Table 15. These include six loci for which the allele ratio was 1, with every allele having a unique amino acid sequence, though these had small allele counts, with a maximum of 4 alleles. Two loci in the virulence functional category were on this list, *relG* ($p < 0.0001$) and *bcpB* ($p < 0.0001$), a toxin

and a probable peroxidoxin, as well as two subunits of the ribosome, *rpsG* ($p < 0.0001$) and *rplN* ($p = 0.0001$). Four other loci had high rates of non-synonymous mutations in Lineage 2, pup ($p < 0.0001$), involved in protein degradation, *sigK* ($p = 0.00099$), involved in the transcription mechanism, *Rv3583c* ($p = 0.00099$), a transcription factor, and *Rv0476* ($p = 0.0001$), with an unknown product.

Table 15. Loci significantly different allele ratio values between Lineage 2 and other lineages.

Locus name	Functional category	Allele ratio outside Lineage 2	Allele ratio within Lineage 2	p-value
pup	Int. metab.	2:7	3:3	9.50E-09
relG	Virulence	4:13	4:4	1.29E-06
bcpB	Virulence	5:16	3:3	3.23E-06
rpsG	rMLST	6:19	2:2	5.90E-06
rplN	rMLST	4:18	2:3	0.0001
Rv0476	Cell wall	4:12	1:1	0.0001
sigK	Information	4:16	5:7	0.00099
Rv3583c	Regulatory	7:20	2:2	0.00099

Six loci were found to differ significantly in their genetic diversity between Lineage 4 and other lineages, with five having lower and one having higher diversity within Lineage 4 than in other lineages, as shown in Table 16. Two loci were involved in the ESX-1 secretion system, *whiB1* ($p = 0.01399$) which is known to regular its expression and *espC* ($p = 0.03731$) which is in the ESX-1 locus. A global transcription regulator that responds to changing oxygen levels, *lsr2* ($p = 0.01398$), is another locus that is strongly conserved in Lineage 4, with only 2 alleles in Lineage compared to 14 alleles in other lineages. A possible toxin, *vapC28* ($p = 01398$) and a locus with an unknown product, *Rv0900* ($p = 01398$), also displayed low diversity. A single locus had much higher diversity within Lineage 4 than in other lineages, with 26 to 9 alleles respectively: *Rv1895* ($p = 0.04335$), a possible dehydrogenase.

Table 16. Loci significantly different values of allelic diversity between Lineage 4 and other lineages

Locus name	Functional category	Average length in base pairs	Unique alleles outside Lineage 4	Unique alleles within Lineage 4	p-value
lsr2	Information	339	14	2	0.01398
Rv0900	Cell wall	153	7	1	0.01398
vapC28	Hypoth.	228	14	2	0.01398
whiB1	Regulatory	255	7	1	0.01399
espC	Cell wall	312	13	2	0.03731
Rv1895	Int. metab.	1155	9	26	0.04335

Nineteen loci had significantly higher allele ratios within Lineage 4 than in other lineages, shown high rates of non-synonymous mutations, as shown in Table 17. This included four loci in the intermediary metabolism functional category, namely *clpP1* ($p = 0.00160$) and *clpP2* ($p = 0.02447$), two subunits of the ClpP protease, responsible for degrading damaged or misfolded proteins (Ollinger et al. 2012), *ppa* ($p = 0.00856$), involved in catalysing pyrosphosphates and *mhuD* ($p < 0.0001$), which produces a heme-degrading protein that is essential for obtaining iron from a host (Graves et al. 2014). Six loci involved in transcription displayed the same dynamic, including *rplN* ($p < 0.0001$), a ribosomal subunit; *Rv1990* and *Rv2640c* (both $p < 0.0001$), probable transcription regulators; *glyS* ($p < 0.0001$), involved in the transcription machinery; *sigH* ($p < 0.0001$), an alternative transcription factor used under oxidative or heat stress; and *rshA* ($p < 0.0001$), whose binds to SigH in order to protect the cell from heat and oxidative stress. *PE20* was the only PE/PPE gene with a high allele ratio in Lineage 4, however it has no known function. This was also the case for *Rv2292c*, *Rv2738c*, *Rv2738c*, *Rv3753c*, *Rv1287* (all $p < 0.0001$), *Rv1116* ($p = 0.00160$) and *Rv0530* ($p = 0.00856$). Lastly, two possible antitoxins, *vapB13* and *vapB17* (both $p < 0.0001$) had significantly higher allele ratios within Lineage 4.

Table 17. Loci significantly different allele ratio values between Lineage 4 and other lineages.

Locus name	Functional category	Average length in base pairs	Allele ratio outside Lineage 4	Allele ratio within Lineage 4	p-value
Rv2292c	Unknown	225	1:5	2:2	1.22281E-37
Rv2738c	Hypoth.	207	2:8	2:2	2.70148E-20
PE20	PE/PPE	300	2:10	11:14	3.03447E-19
vapB17	Virulence	228	1:5	3:4	9.87725E-17
vapB13	Virulence	264	2:7	4:4	1.76768E-13
rplN	rMLST	369	2:13	4:8	1.53495E-10
Rv3753c	Hypoth.	501	2:11	5:9	1.86963E-08
Rv1287	Hypoth.	486	2:8	6:8	6.88177E-08
Rv1990c	Regulatory	342	2:7	5:6	4.46361E-07
glyS	Information	1392	6:24	8:11	5.25997E-07
sigH	Information	651	6:17	12:12	2.66204E-06
Rv2640c	Regulatory	360	4:15	6:8	4.17481E-06
mhuD	Int. metab.	318	2:8	7:10	5.36951E-06
rshA	Information	306	3:8	6:6	7.64756E-05
Rv1116	Hypoth.	186	2:5	4:4	0.00160
clpP1	Int. metab.	603	3:15	8:16	0.00160
ppa	Int. metab.	489	5:20	6:10	0.00856
Rv0530	Hypoth.	162	1:4	3:5	0.00856
clpP2	Int. metab.	645	3:11	7:11	0.02447

Homoplasy

Convergent evolution can be observed in some loci, when alleles appear in multiple lineages. A majority of loci, 3,594 out of 3,795 (94.7%), included up to two alleles that appeared in both isolates assigned to Lineage 2 and to Lineage 4. Nineteen loci displayed high levels of homoplasy, including four loci known to be associated with drug resistance, namely *pncA* (25/160 alleles in both lineages), *rpoA* (5/28 alleles), *inhA* (6/23 alleles) and *ahpC* (4/21 alleles). Three loci with ESAT-6 like protein, *esxN* (10/59 alleles), *esxV* (5/26 alleles) and *esxJ* (8/16 alleles) display high levels of homoplasy. A further two loci are placed in the cell wall functional category, but have no known product, namely *Rv1312* (4/13 alleles) and *lppP* (4/21). Only one locus in the PE/PPE category appeared in this list, *PPE13* (6/33 alleles); it also had no known function. Virulence-associated loci were present, as *kasB*, whose product can regulate virulence via involvement in the synthesis of mycolic acids (Vilcheze et al. 2014) and *ripD* (5/28 alleles) which contributes to the cell wall's defensive properties . The remaining seven loci are hypothetical: *Rv0740* (4/14 alleles), *Rv1778c* (5/10), *Rv2307D* (5/10), *Rv2313c* (4/25), *Rv2722* (4/11), *Rv2983* (4/21) and *Rv3172c* (4/13).

Table 18. Loci with high levels of homoplasy, as determined by a high proportion of alleles shared between Lineage 2 and Lineage 4.

Locus name	Functional category	Total unique alleles in both lineages	Alleles observed in both Lineage 2 and Lineage 4
pncA	Int. metab.	160	25
Rv2313c	Hypoth.	25	4
esxN	Cell wall	59	10
Rv1566c	Virulence	28	5
rpoA	Information	28	5
kasB	Lipid	22	4
PPE13	PE/PPE	33	6
lppP	Cell wall	21	4
ahpC	Virulence	21	4
Rv2983	Hypoth.	21	4
esxV	Cell wall	26	5
inhA	Lipid	23	6
Rv0740	Hypoth.	14	4
Rv1312	Cell wall	13	4
Rv3172c	Hypoth.	13	4
Rv2722	Hypoth.	11	4
Rv1778c	Hypoth.	10	5
Rv2307D	Hypoth.	10	5
esxJ	Cell wall	16	8

Discussion

Allele designations

The MTBC have low genetic diversity and evolve purely though clonal descent (*M. canettii* excepted). These features make the MTBC the ideal model organism for large-scale genomic investigations such as this, as they are well suited to automated curation. Over twenty million allele designations were made for this dataset, using a combination of automation and manual curation that creates a high-quality annotation of a large dataset, now available for further analysis as well. Full records are kept of every stage of curation, enabling future work to continue seamlessly.

The curation of allele designations also presented a preliminary indication of the low levels of genetic diversity present in the species complex. Nearly three quarters of all allele designations made were exact matches of the alleles present in the reference genome.

Lineage assignment and phylogenetic structure

Despite the use of a scheme including only the most conserved loci in the entire *Mycobacterium* genus, core genome profiles were successful in distinguishing lineage, and occasionally sub-lineage, clusters within the MTBC. The use of core genome profiles, which combine designations of 749 loci into one integer identifier, provided a method by which most isolates could be uniquely referenced and the most closely related isolates, likely collected from the same outbreak or transmission chain, could be grouped. These profiles demonstrate the power of combining hundreds of loci, which individually would provide negligibly low amounts of resolution, into a profile with ample resolution to distinguish isolates. Around 6% of isolates were not assigned a cgST, despite profiles having a tolerance of up to 50 missing loci. This could be improved by the creation of a MTBC-specific core genome with loci specially chosen for their

high rates of presence as full coding sequences in assembled MTBC genomes. However, this would negate the utility of having the entire *Mycobacterium* genus characterised using a common core genome scheme. While isolates from other species in the genus were not included in this analysis beyond the creation of core genome scheme, they could readily be placed in a genus-wide phylogeny using this scheme.

Single-linkage clustering of these profiles succeeded in forming isolates groups that matched known existing lineages. Some lineages included multiple groups, though this may be due to lack of representation within that lineage. Single-linkage clustering merges together groups in which even a single isolate is below the set threshold of similarity, so fuller sampling leads to fewer groups given the same diversity in the source population. A similar effect may happen to the smallest groups. Groups of three or fewer members could in most cases not be assigned to a lineage, but with fuller sampling they may merge with the larger groups. Alternatively, a more tolerant threshold may be able to group them. The clustering of all three modern lineages isolates together in a single classification scheme group at the more tolerant threshold of 120 allelic mismatches underscores their recent common ancestry. The presence of distinct sub-lineage clusters within the stricter classification scheme, of 75 allelic mismatches, points to the diversity existing within lineages, with known differences in pathobiology observed between sub-lineages (Stucki et al. 2016).

The phylogenetic relationships determined by the neighbour-joining method from cgMLST allele designations are congruent with the consensus phylogeny for the species complex as determined by a variety of molecular methods (Niemann et al. 2016). This is to be expected, as all of these methods rely on DNA sequences, whether they are assessed by allele mismatches, SNP differences, or by the deletion of "regions of genomic difference" (RDs), because descent is strictly clonal in nearly all species in the complex. In a fully clonally evolving population, any

level of resolution that can sufficiently distinguish strains will reflect the true phylogenetic relationship, except in cases of strong selection leading to congruent evolution.

The number of allele mismatches between pairs of isolates is not evenly distributed over its range. Multiple distinct peaks in the distribution appear as the result of strong population structuring within the MTBC, with peaks corresponding to the average allele mismatches at within lineage, within the "modern" lineages (Lineages 2, 3 and 4) and within species levels. The varying levels of diversity within each lineage were further elucidated by the maximum number of mismatches between members belonging to the same lineage. Unsurprisingly, this is congruent with previous results. The modern lineages show restricted levels of diversity, while *M canettii*, the outgroup and possible environmental reservoir ancestor, displays the most diversity. In fact, isolates of *M. canettii* were as genetically distant from one another as they were from other members of the species complex, with the exception of isolates that were nearly identical, presumably as a consequence of being repeated sequencing or assembly runs of the same lab isolate. This level of diversity is even more remarkable considering *M. cannetti* is geographically limited to the Horn of Africa, whereas the modern lineages are found worldwide. *M. canettii* has for this reason been speculated to be the ancestral population of the species complex, or possibly be exchanging genetic material with a large and diverse reservoir (Coscolla and Gagneux 2014).

Diversity and selection

While close to four thousand loci were included in this analysis, only a few dozens were found to differ from the average in any of the traits investigated.

Loci varied significantly in genetic diversity. The most diverse locus having had 40 times more alleles per nucleotide length than the least diverse, and the proportion of variable sites in an alignment varied from less than 1% to more than 43%.

The least diverse loci were in two cases known to occasionally undergo gene fusion, causing them to have allele designations in fewer isolates. In other cases, they were hypothetical conserved loci, with no known function. Loci with high allele diversity, in contrast, were often alleles with known functions in virulence. The included two loci whose mutations may confer drug resistance, *gid* and *pncA*, the latter of which is also notable for displaying high levels of homoplasy, and three loci related to the ESAT-6 like family of genes, including *esxN* and *esxO*, and *whiB6*. PE/PPE genes, known for a variable number of the repeats of the sequences with the "PE" and "PPE" motifs, were among the loci with the most variables sites in their alignments. These loci had high genetic diversity as measured by alignment-based measures but moderate levels under allele-based measures, highlighting a useful contrast between these measures, notwithstanding their close correlation in the case of alleles differing at single nucleotide polymorphisms only. PE and PPE genes are unique to *Mycobacterium* and make up around 10% of the *M. tuberculosis* genome. Few of these genes have known functions, but their many recently duplications and the combination of conserved structure and highly variable motif repeats may point to a function in the production of antigenic variation and therefore virulence (Akhter et al. 2012). A resuscitation promotion factor, *rcfA*, also displayed high levels of diversity. The protein factor coded for by this locus is known to be essential for virulence in an animal model. Changes in the make-up of this protein may cause changes in the development of tuberculosis disease within an infected patient (Rosser et al. 2017). The conditions under which the infections leaves dormancy may be especially important for the adaptation of tuberculosis to different human populations. Human communities vary widely in their density. A strongly virulent strain of *M. tuberculosis* in a small isolate village may wipe the entire population before it has a chance to spread to new communities. In contrast, the same strain may be wildly successful in a large city, where a successful transmission after fleeting contacts may allow it to spread very quickly.

Loci with high measures of selection, in the form of a high ratio of proportions of variable sites, had various functions. Some were involved in essential cell functions such as arsenic transport, arginine biosynthesis and cholesterol metabolism, while others were known to directly affect the cytokine response in the host. This measure pointed to many loci with unknown functions, perhaps highlighting a fruitful area of future research. Loci found here to appear conserved in the genome may have essential functions, possibly linked to the highly specialised lifecycle of *M. tuberculosis*. Loci found to be more diverse than the average may have interactions with the host immune system that increase their pressure to diversify.

Many loci displaying signals of strong purifying selection, as a low rate of non-synonymous mutations to synonymous mutations, were crucial to the survival of the *M. tuberculosis* cell. A locus whose product is an essential part of the folate biosynthesis pathway, *folE*, was the most remarkably conserved, with significantly low values for allele ratio and proportion ratio. This indicates it is under extremely strong purifying selection, where even a small deviation from its wild type causes the death of the cell in which the mutation occurs. This may point to this locus as a good target for DNA-based diagnostic tests. Other essential loci included *clpX*, which produces the Clp-ATPase, and groEL2. The latter produces Cpn60.2, a chaperonin that is essential for *M. tuberculosis* to survive in the host. As *M. tuberculosis* is an obligate pathogen, this protein is both essential for growth and involved in virulence of the bacteria.

Lineage-specific effects

Lineage 2

Only one locus was found to have much higher relative diversity within Lineage 2 than in other lineages, a ferrodoxin that transports electrons. This could potentially point to an adaptation to the poor availability of iron in the host. Five loci had relatively lower diversity within Lineage 2, however these had no known function or simple essential metabolism

functions. Despite early reports that the successful propagation of Lineage 2 could be attributed to a high mutation rate, including perhaps mutational 'bursts' in response to stress, later studies have not found evidence of this effect, and causing the success of this lineage to remain an open question (Merker et al. 2013).

Lineage 4

In both Lineage 2 and Lineage 4, no loci appeared to have signals of strong purifying pressure. Instead, an abundance of loci with high rates of synonymous mutations was observed. This may point to the accumulation of deleterious mutations in essential loci. Due to the fully clonal nature of MTBC evolution, strong positive selection on a single locus, for example when a mutation confers drug resistance, can allow slightly deleterious mutations to "hitchhike" them (Smith et al. 2006). These high rates of synonymous mutations in essential genes points to this effect occurring in the rapidly expanding Lineages 2 and 4.

Lineages 2 and 4 are the most widespread and most virulent, causing a majority of the disease burden of tuberculosis today (Stucki et al. 2016). Given the fully clonal nature of both lineages, none or perhaps a single ancestral allele would be expected to be found in both these lineages at each locus**.** Instead, the most common number of shared alleles was two. This may reflect the high prevalence of mixed infections, which can be hard to distinguish, especially when composed of closely related strains (Shamputa et al. 2006). MTBC strains are particularly vulnerable to this, due to their slow-growing nature and low levels of diversity. The loci displaying the highest levels of homoplasy did not appear to be random, however, as they included four drug resistance associated loci. Homoplasy has previously been reported as occurring at high rates in drug resistance associated genes (Eldholm et al. 2014).

The ESAT-6 protein family and its associated secretion systems showed distinct differences from the average of the genome in various measures. There are five ESX (ESAT secretion)

systems in the *M. tuberculosis* genome, including ESX-1, whose deletion is responsible for the attenuation of virulence that allows the BCG strain to be used as a human vaccine. Three loci, *esxO*, *esxN*, and *whiB6*, displayed high allelic diversity, while *whiB1* and *espC* displayed high allele ratios and are present in the ESX-1 locus. Lastly, three potentially secreted proteins in the ESAT-6 family, *esxN*, *esxV* and *esxJ*, displayed high levels of homoplasy. While ESX-1 has been extensively characterised, the remaining four secretions systems have effects on virulence but no clear function (Groschel et al. 2016). The high levels of diversity and diversifying selection shown by these loci points to possibly important functions that remain to be discovered.

Further work

While MLST databases such as the one used here are portable, curation is an on-going process. Adding novel locus definitions and alleles, and new isolates, improve the completeness of the database, and can be completed without the need to edit any previous completed curation. The addition of genomes from rarer lineages, especially finished genomes, can add to the quality of the dataset. Finished genomes ensure that missing allele designations are caused by legitimate deletions (or gene fusions), rather than problems with genome assemblies. Likewise, more genomes assembles from larger read sequences can contribute to the quality of loci designations by more often including complete copies of loci with repeated sequences.

While these results are possible from a database composed purely of MTBC genetic sequences, confirmation of the genetic basis for virulence requires more extensive data. This is ideally accomplished by the comparison of isolates with known varying virulence, including data about the human patient, his or her environment, and clinical outcomes.

Supplemental files

S1 Table with allele designations for 6,679 MTBC isolates and 3,860 loci

S2 Locus information table, including gene name, functional category, Rv number and essentiality measure by Griffin et al.

S3 Isolate information table, including lineage assignments and assembly statistics

S4 Dissimilarity matrix for 6,679 MTBC isolates

S5 GUAVA Results Table for 6,679 MTBC isolates and 3,860 loci

Chapter III ***References***

Akhter Y, Ehebauer MT, Mukhopadhyay S, Hasnain SE (2012) The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: Perhaps more? Biochimie 94:110–116. doi: 10.1016/j.biochi.2011.09.026

Bazet Lyonnet B, Diacovich L, Cabruja M, et al (2014) Pleiotropic effect of AccD5 and AccE5 depletion in acyl-coenzyme A carboxylase activity and in lipid biosynthesis in mycobacteria. PloS One 9:e99853. doi: 10.1371/journal.pone.0099853

Bhatia AL, Csillag A, Mitchinson DA, et al (1961) The virulence in the guinea-pig of tubercle bacilli isolated before treatment from South Indian patients with pulmonary tuberculosis. 2. Comparison with virulence of tubercle bacilli from British patients. Bull World Health Organ 25:313–322.

Billig S, Schneefeld M, Huber C, et al (2017) Lactate oxidation facilitates growth of Mycobacterium tuberculosis in human macrophages. Sci Rep 7:6484. doi: 10.1038/s41598-017-05916-7

Both D, Steiner EM, Izumi A, et al (2014) RipD (Rv1566c) from Mycobacterium tuberculosis: adaptation of an NlpC/p60 domain to a non-catalytic peptidoglycan-binding function. Biochem J 457:33–41. doi: 10.1042/BJ20131227

Braden CR, Morlock GP, Woodley CL, et al (2001) Simultaneous infection with multiple strains of Mycobacterium tuberculosis. Clin Infect Dis 33:e42–7. doi: 10.1086/322635

Camus JC, Pryor MJ, Medigue C, Cole ST (2002) Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiol Immunol 148:2967–2973.

Chen Z, Hu Y, Cumming BM, et al (2016) Mycobacterial WhiB6 Differentially Regulates ESX-1 and the Dos Regulon to Modulate Granuloma Formation and Virulence in Zebrafish. Cell Rep 16:2512–2524. doi: 10.1016/j.celrep.2016.07.080

Cohen KA, Abeel T, Manson McGuire A, et al (2015) Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. PLoS Med 12:e1001880. doi: 10.1371/journal.pmed.1001880

Cole ST, Brosch R, Parkhill J, et al (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. 393:537–. doi: 10.1038/31159

Comas I, Coscolla M, Luo T, et al (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet 45:1176–1182. doi: 10.1038/ng.2744

Comas I, Homolka S, Niemann S, Gagneux S (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PloS One 4:e7815–e7815. doi: 10.1371/journal.pone.0007815

Coscolla M, Gagneux S (2014) Consequences of genomic diversity in Mycobacterium tuberculosis. Semin Immunol 26:431–444. doi: 10.1016/j.smim.2014.09.012

Daim S, Kawamura I, Tsuchiya K, et al (2011) Expression of the Mycobacterium tuberculosis PPE37 protein in Mycobacterium smegmatis induces low tumour necrosis factor alpha and interleukin 6 production in murine macrophages. J Med Microbiol 60:582–591. doi: 10.1099/jmm.0.026047-0

Donoghue HD (2009) Human tuberculosis - an ancient disease, as elucidated by ancient microbial biomolecules. Microbes and Infection 11:1156–1162. doi: 10.1016/j.micinf.2009.08.008

Eldholm V, Norheim G, Lippe von der B, et al (2014) Evolution of extensively drug-resistant Mycobacterium tuberculosis from a susceptible ancestor in a single patient. Genome Biol 15:490. doi: 10.1186/s13059-014-0490-3

Fenner L, Egger M, Bodmer T, et al (2013) HIV infection disrupts the sympatric host-pathogen relationship in human tuberculosis. PLoS Genetics 9:e1003318. doi: 10.1371/journal.pgen.1003318

Gagneux S (2013) Genetic diversity in Mycobacterium tuberculosis. Curr Top Microbiol Immunol 374:1–25. doi: 10.1007/82_2013_329

Gagneux S (2012) Host-pathogen coevolution in human tuberculosis. Phil Trans R Soc B 367:850–859. doi: 10.1098/rstb.2011.0316

Gagneux S, Small PM (2007) Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. Lancet Infect Dis 7:328–337. doi: 10.1016/S1473-3099 (07)70108-1

Gandhi NR, Nunn P, Dheda K, et al (2010) Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. Lancet 375:1830–1843. doi: 10.1016/S0140-6736 (10)60410-2

Graves AB, Morse RP, Chao A, et al (2014) Crystallographic and Spectroscopic Insights into Heme Degradation by Mycobacterium tuberculosis MhuD. Inorg Chem 53:5931–5940. doi: 10.1021/ic500033b

Griffin JE, Gawronski JD, DeJesus MA, et al (2011) High-Resolution Phenotypic Profiling Defines Genes Essential for Mycobacterial Growth and Cholesterol Catabolism. PLoS Pathogens. doi: 10.1371/journal.ppat.1002251

Groschel MI, Sayes F, Simeone R, et al (2016) ESX secretion systems: mycobacterial evolution to counter host immunity. Nature Publishing Group 14:677–691. doi: 10.1038/nrmicro.2016.131

Hershberg R, Lipatov M, Small PM, et al (2008) High Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human Demography. PLoS Biol 6:e311. doi: 10.1371/journal.pbio.0060311

Hirsh AE, Tsolaki AG, DeRiemer K, et al (2004) Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proc Natl Acad Sci USA 101:4871–4876. doi: 10.1073/pnas.0305627101

Höner Zu Bentrup K, Miczak A, Swenson DL, Russell DG (1999) Characterization of activity and expression of isocitrate lyase in Mycobacterium avium and Mycobacterium tuberculosis. J Bacteriol 181:7161–7167.

Joseph S, Yuen A, Singh V, Hmama Z (2017) Mycobacterium tuberculosis Cpn60.2 (GroEL2) blocks macrophage apoptosis via interaction with mitochondrial mortalin. Biol Open 6:481–488. doi: 10.1242/bio.023119

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. doi: 10.1093/molbev/mst010

Kohl TA, Diel R, Harmsen D, et al (2014) Whole genome based Mycobacterium tuberculosis surveillance: A standardized, portable and expandable approach. J Clin Microbiol 52:JCM.00567–14–2486. doi: 10.1128/JCM.00567-14

Lew JM, Kapopoulou A, Jones LM, Cole ST (2011) TubercuList--10 years after. Tuberculosis (Edinb) 91:1–7. doi: 10.1016/j.tube.2010.09.008

Maiden MCJ, Jolley KA (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595. doi: 10.1186/1471-2105-11-595

Maiden MCJ, Jolley KA, Parkhill J, et al (2012) A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus Neisseria. Microbiol 158:1570–1580. doi: 10.1099/mic.0.056077-0

Maiden MCJ, van Rensburg MJJ, Bray JE, et al (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:728–736. doi: 10.1038/nrmicro3093

Meena LS, Meena J (2016) Cloning and characterization of a novel PE_PGRS60 protein (Rv3652) of Mycobacterium tuberculosis H37 Rv exhibit fibronectin-binding property. Biotechnol Appl Biochem 63:525–531. doi: 10.1002/bab.1411

Merker M, Kohl TA, Roetzer A, et al (2013) Whole genome sequencing reveals complex evolution patterns of multidrug-resistant Mycobacterium tuberculosis Beijing strains in patients. PloS One 8:e82551. doi: 10.1371/journal.pone.0082551

Niemann S, Merker M, Kohl T, Supply P (2016) Impact of Genetic Diversity on the Biology of Mycobacterium tuberculosis Complex Strains. doi: 10.1128/microbiolspec.TBTB2-0022-2016

Ollinger J, O'Malley T, Kesicki EA, et al (2012) Validation of the Essential ClpP Protease in Mycobacterium tuberculosis as a Novel Drug Target. J Bacteriol 194:663–668. doi: 10.1128/JB.06142-11

Rathor N, Garima K, Sharma NK, et al (2016) Expression profile of mce4 operon of Mycobacterium tuberculosis following environmental stress. Int J Mycobacteriol 5:328–332. doi: 10.1016/j.ijmyco.2016.08.004

Rosser A, Stover C, Pareek M, Mukamolova GV (2017) Resuscitation-promoting factors are important determinants of the pathophysiology in Mycobacterium tuberculosis infection. Crit Rev Microbiol 43:621–630. doi: 10.1080/1040841X.2017.1283485

Shamputa IC, Jugheli L, Sadradze N, et al (2006) Mixed infection and clonal representativeness of a single sputum sample in tuberculosis patients from a penitentiary hospital in Georgia. Respir Res 7:99. doi: 10.1186/1465-9921-7-99

Sherlin D, Anishetty S (2015) Mechanistic insights from molecular dynamic simulation of Rv0045c esterase in Mycobacterium tuberculosis. J Mol Model 21:90. doi: 10.1007/s00894-015-2630-4

Sinsimer D, Huet G, Manca C, et al (2008) The phenolic glycolipid of Mycobacterium tuberculosis differentially modulates the early host cytokine response but does not in itself confer hypervirulence. Infect Immun 76:3027–3036. doi: 10.1128/IAI.01663-07

Smith NH, Gordon SV, la Rua-Domenech de R, et al (2006) Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat Rev Microbiol 4:670–681. doi: 10.1038/nrmicro1472

Stucki D, Brites D, Jeljeli L, et al (2016) Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nature Publishing Group 48:1535–1543. doi: 10.1038/ng.3704

Vilcheze C, Molle V, Carrere-Kremer S, et al (2014) Phosphorylation of KasB Regulates Virulence and Acid-Fastness in Mycobacterium tuberculosis. PLoS Pathogens. doi: 10.1371/journal.ppat.1004115

World Health Organization (2016) Global Tuberculosis Report 2016.

Chapter IV
Patterns of genetic diversity and evolutionary selection in the core genomes of four bacterial species

Sofia Hauck, James E. Bray, Keith A. Jolley, Martin C. J. Maiden

Department of Zoology, University of Oxford, Oxford, United Kingdom

Abstract

Background: Correct identification of bacterial pathogens is key for treatment and prevention. The speed, low cost and high resolution of characterisation possible with whole genome sequencing has led it to its widespread use, and to the accumulation hundreds of thousands of whole genome sequences for bacterial species. These can be used to examine the population genetics of the species which have been samples.

Methods: Four bacterial pathogen species were chosen: *Campylobacter jejuni*, *Mycobacterium tuberculosis*, *Neisseria meningitis* and *Staphylococcus aureus*. Universal ribosomal gene sequences were obtained for 73,145 isolates. Species-specific core genome sequences were obtained for 49,859 isolates. Allele discovery curves were plotted to estimate the completeness of sampling, and the levels of diversity and recombination of each species. Allele-based methods were used to quantify the signals of diversity and selection at the ribosomal loci, and signals of selection at core loci.

Results: *C. jejuni* and *S. aureus* showed high levels of sequence diversity, while *N. meningitidis* and to an even greater degree, *M. tuberculosis*, showed restricted levels of sequence variation. *C. jejuni* and *N. meninigitis* displayed high levels of recombination in their core genes, while *M. tuberculosis* and *S. aureus* appear to have core genomes that evolve clonally. Ribosomal loci were shown to have no pattern in their diversity across species, and only a slight effect the strength of selection, with the smaller subunit being more strongly conserved.

Conclusion: Ribosomal genes remain a useful locus scheme, and reflect the patterns of selection and diversity seen elsewhere in the core genome. We confirm the variation in genetic diversity and selection strength for four common bacterial pathogens is as predicted by smaller-scale studies, using with the largest data set analysed yet.

*Chapter IV: **Table of Contents***

Abstract **1**

Background **4**

Methods **7**

Isolates **7**

Table 1. Numbers of isolate sequences 8

Loci **8**

Ribosomal MLST 8

Core genome 9

Allele-based measures of diversity **10**

Statistical tests **10**

Results **12**

Table 2. Profiles and total alleles in rMLST sequences.....12

Figures 1a and 1b. Discovery curves of total alleles in rMLST loci found in four species. .13

Table 3. Allelic diversity in rMLST genes, measured in alleles per nucleotide. 16

Figure 2. Allelic diversity in the rMLST loci over four species. 16

Figure 3. Allelic diversity in the 53 rMLST loci over four species. 17

Table 4. Selection measures for rMLST loci in four species. 18

Figure 4. Allele ratios in the rMLST loci over four species. 19

Figure 5. A heatmap shows allelic diversity, measured as ratio of unique alleles sequences after and before translation.....20

Table 5. Selection measures for core genome loci in four species. 21

Figure 6. Allele ratios in the core genome loci of four species. 22

Discussion **23**

References **31**

Background

Accurate and timely identification of bacterial pathogens is crucial for the treatment of patients and for the management of outbreaks. Basing identification of bacterial isolates on molecular characteristics, such as enzymatic activity or the presence of specific nucleic acids, has vastly improved the quality of diagnostics over purely phenotypic techniques (Sintchenko et al. 2007). However, early molecular techniques were cumbersome, requiring expensive and difficult-to-acquire reagents, and large amounts of laboratory time per sample. Worse yet, these techniques often had poorly reproducible outputs, making it impossible to compare across laboratories and precluding the sharing of data at a multinational level (Salipante et al. 2015).

Multi-locus sequence typing, also known as MLST, was developed partly as an answer to this portability problem. MLST characterises isolates based on their genetic sequences at specific loci, typically housekeeping genes or fragments thereof, using arbitrary integers to refer to specific versions of these loci, termed alleles. Comparing isolates based on their number of matching or mismatching alleles at multiple loci enables rapid and accurate indexing of isolate collections (Maiden et al. 1998). Horizontal gene transfer, also known as HGT, occurs frequently in most bacterial species, and is accommodated for in the MLST model by the use of multiple loci spread throughout the genome. This ensures a fuller view of the genome than is possible in a method that relies on a single locus, which may change considerably if subject to a recombination event, despite little change to the genome at large. Compared to previous molecular techniques, MLST represented a dramatic improvement in resolution, reproducibility and portability (Maiden et al. 2013).

MLST schemes use six to eight loci, a number sufficient for typing within a species. However, the advent of next-generation whole genome sequencing (WGS) has opened the possibility of schemes with far more loci. The extraction of alleles from a draft genome sequence is trivial compared to the resources required to isolate colonies in the lab, prepare

DNA for sequencing, assemble the short reads produced by WGS, and so on. MLST schemes with less than ten loci had limited utility due to their limited resolution and limited range. With such few loci, MLST schemes must be specific to closely related isolates, typically within one species or genus, which precludes multi-species analyses. Examining only a handful of loci also limits what genetic diversity can be uncovered. The MLST scheme used for *N. meningitidis*, for example, indexes around 0.15% of its genome (Jolley and Maiden 2013).

The ribosomal MLST scheme, often abbreviated to rMLST, was created to address these limitations. It takes advantage of the universal presence of the eponymous ribosomes, which are found in all bacteria. Each ribosome is composed of a small and a large subunit, both of which contain many proteins. The combination of a set of loci universal to all bacteria and the resolution afforded by over fifty loci distributed throughout the genome created a scheme capable of identifying isolates at every taxonomic level from domain to sub-species strain (Jolley et al. 2012).

The PubMLST database (https://pubmlst.org), started in 2010, utilises the Bacterial Isolate Genome Sequence Database, also known as BIGSdb, software to maintain both the universal typing scheme, rMLST, and hundreds of more specific databases (Maiden and Jolley 2010). The rMLST database especially has seen phenomenal growth, and holds as of the time of writing 219,727 isolate records for 7,328 species. There is a strong skew in isolate distribution by species however, as only 83 species (1.13%) are represented by more than 100 isolates each in the database. For the rMLST scheme alone, PubMLST contains 860,427 unique allele sequences.

The rMLST scheme has been used primarily for the classification and typing of isolates. The loci in the rMLST scheme, however, are like all other genes in a bacterial genome in that they are subject to the processes of evolution. Mutations that occur randomly with respect to gene function during replication or directly through DNA damage can affect the genetic sequences of the ribosomal genes. The different structures of the ribosomal proteins that can

results from these mutations are capable of influencing the fitness of the bacteria in which they are found. According to Ohta's Nearly Neutral Theory, most mutations have only a small effect on the fitness of the organism, and these are mostly slightly deleterious effects, as genomes are highly optimised and random changes are more likely to disrupt it than provide a fitness improvement (Ohta 1992; Ohta 2011). Because mutations are typically random with respect to gene function, they occur at similar rates in all parts of the genome. Evolutionary selection acts on those mutations, increasing or decreasing their frequency through the differential survival and reproduction of the organisms that carry them. Differences in the genetic diversity of loci therefore can be marks left on the genome by evolutionary forces.

The large collection that has been amassed for the benefit of improved identification of pathogens can therefore be used for a second purpose, of uncovering the evolutionary forces and mechanisms at play in various bacterial species. It also allows for a review of the ribosomal loci. These loci were chosen for to make up the universal scheme due to their shared function. As of yet, no large-scale multi-species analysis has investigated the variation between these genes. Here we used this data set to investigate the signals of diversity and selection across bacterial species, and between the loci in the rMLST scheme.

Methods

Isolates

Isolate records, including allele designations, and the genetic sequences referred to by these designations, were collected from databases hosted by PubMLST (https://pubmlst.org/).

Species were chosen on the following criteria:

a) the species is within the Bacteria domain,

b) a whole or core genome scheme for the species or its genus is available,

c) more than 10,000 isolates of that species exist in the universal ribosomal MLST isolates database,

d) more than 7,000 isolates containing whole genome sequences exist for that species.

These criteria fit four species, namely *Campylobacter jejuni*, *Mycobacterium tuberculosis*, *Neisseria meningitidis* and *Staphylococcus aureus*. The number of isolates records for these species in the "Ribosomal MLST genome database" ranged from 11,840 *N. meningitidis* isolates to 33,150 *S. aureus* isolates, as shown in Table 1.

Species or genus-specific databases existed for each of these four species, containing whole genome sequences. Whole genome sequences of *C. jejuni* were obtained from the "Campylobacter jejuni/coli PubMLST database" and limited to *C. jejuni* isolate records with a ribosomal sequence type (rST) that was also present in the "Ribosomal MLST genome database" *C. jejuni* isolate records. Likewise, *N. meningitidis* isolates were collected from the "Neisseria PubMLST database" and limited to those with rSTs that also existed within the "Ribosomal MLST genome database" *N. meningitis* records. Whole genomes for *M. tuberculosis* were obtained from the "Mycobacteria genomes database" and limited to isolate records using the same draft assemblies or finished genome sequences as Ribosomal MLST *M. tuberculosis* records, as defined by isolate record IDs in the "Ribosomal MLST

genome database" that matched rMLST database accession IDs in the "Mycobacteria genomes database". Whole genomes of *S. aureus* were obtained in a similar way, with matching IDs and ribosomal accession IDs in the "Staphylococcus aureus isolates database". The number of whole genome sequences was generally less than the number of ribosomal MLST sequences, ranging from 7,509 *M. tuberculosis* sequences to 25,512 *S. aureus* sequences.

Table 1. Numbers of isolate sequences

Species name	Isolates with ribosomal sequences	Isolates with whole genome sequences
Campylobacter jejuni	15,654	8,781
Mycobacterium tuberculosis	12,501	7,509
Neisseria meningitidis	11,840	8,057
Staphylococcus aureus	33,150	25,512
Overall	73,145	49,859

Loci

Ribosomal MLST

The universal bacteria typing scheme named Ribosomal MLST (or rMLST) is composed of 53 loci. These loci code for the protein subunits of the ribosome, and include 21 proteins in the small, also called 30S, subunit and 31 proteins in large, also called 50S, subunit. Not every locus is present in every species, and some species include gene duplication of rMLST loci, causing a pair of paralogous loci to exist in the genome. In the four species included in this analysis, only two species have rMLST locus deletions: *M. tuberculosis* has no rpsU locus, and *N. meningitidis* has no rpmD locus. Paralogous loci are more common, occurring in three species. *N. meningitidis* has paralogues of rpmE and rpmJ; *S. aureus* genomes contains paralogues of rpsN and rpmG; and lastly *M. tuberculosis* has four pairs of paralogues, for rpsN, rpsR, rpmB and rpmG.

Core genome

Core genomes were determined for each species as loci in core or whole genome schemes with allele designations in over 90% of isolates with whole genome sequences. Existing locus schemes for the four species in this analysis varied in their scope and curation efforts. For *M. tuberculosis*, a whole genome scheme derived from the second revision of the H37Rv reference genome (Köser et al. 2012) exists in PubMLST. It excludes non-paralogous ribosomal genes, non-protein genes and genes that do not form a complete coding sequence (detailed in Chapter III). The existing locus scheme for *S. aureus* is based on a core genome sensu stricto for the species, with every locus present in a collection of 181 isolates study (Meric et al. 2015). The *S. aureus* database is not manually curated, relying solely on the automatic curation features of BIGSdb. The *Neisseria* database, in contrast, is heavily curated by multiple people, and has been the basis for multiple investigations of *Neisseria* genomics, especially in outbreak contexts, and includes a core genome scheme for the *Neisseria* genus (Jolley and Maiden 2013; Bratcher et al. 2014; Maiden and Harrison 2016). The *Campylobacter* database is similar, with detailed curation. It contains hundreds of isolates from disease cases in Oxfordshire but also isolates from various other countries and animal sources. It also contains a core genome scheme for the *Campylobacter* species *C. coli* and *C. jejuni* (Cody et al. 2013; Jansen van Rensburg et al. 2016; Cody et al. 2017).

Only loci with designations in more than 90% of whole genome isolates were determined as forming part of the core genomes in this analysis. This stringent cutoff limited the core genome to 1,375 loci out of a possible 1,934 (71.1%) for *C. jejuni*, 3,172 out of 3,942 (80.5%) for *M. tuberculosis*, 1,546 out of 2,692 (57.4%) *N. meningitidis* and 1,965 out of 2,209 (89.0%) for *S. aureus*. This criterion can reflect the prevalence of loci in the genomes of isolates from the relevant species, but can also be affected by the quality of whole genome sequencing, the quality of genome assembly in the case of draft genomes, and the quality of allele calling methods.

Allele-based measures of diversity

Allelic diversity was measured using the GUAVA (detailed in Chapter 2) application. This measure is based the unique allele sequences of each locus, measuring their total number and their average length in base pairs. The former is divided by the latter, giving a measure of alleles per nucleotide, analogous to single nucleotide polymorphism (SNP) density. It differs from SNP density in that it measures the entire locus at once, and in that insertions and deletions are also captured as genetic differences.

Allele ratio was also measured with GUAVA. For each locus, the total number of unique sequences before and after translation is counted. The ratio of these counts is taken, giving a measure of the rate of synonymous to non-synonymous mutations, analogous to dN/dS ratio or KaKs ratio.

Statistical tests

To calculate whether the differences between the allelic diversity or allele ratio values for each species differed across the rMLST loci, one-way analysis of variance (ANOVA) tests were performed. The locus was included as a blocking factor, and paralogous loci were excluded. The pairwise differences between the species, as factors of the ANOVA test, were assessed using the Tukey Honest Significant Difference (HSD) test.

Mann-Whitney tests were used to compare the ranking of loci over their allelic diversity and allele ratio values across the four species. Each pairwise comparison between the four species was assessed. Only loci that were neither paralogous nor absent in any of the four species were included (45 in total).

Differences in allelic diversity and allele ratio between the loci of the small and large subunits of the ribosome were assessed using Student's t-test.

Results

Species varied in the number of ribosomal sequence types (rSTs) and the total number of unique allele sequences found in their isolates, as shown in Table 2. This variance could not be fully explained by to the number of isolate records for each species, as *M. tuberculosis* had roughly a third as many ribosomal sequence types as *N. meningitidis* (1,118 and 3,267 respectively) despite a comparable number of isolate records for both species, with *N. meningitidis* having fewer between the two (12,501 and 11,840 and respectively). This contrast was less stark in the total number of rMLST alleles discovered, with *N. meningitidis* still having a greater number, but less than double the total count found in *M. tuberculosis* isolates (3,436 to 2,006 alleles).

The rate at which new profiles of rMLST loci, also known as ribosomal sequence types or rSTs, are discovered is not uniform between the four species. The total number of *S. aureus* and *C. jejuni* rMLST alleles discovered was comparable, at 6,226 and 6,514 respectively, however these alleles were found in 49.4% more combinations in *S. aureus* than in *C. jejuni* (4,524 to 3,028 rSTs).

The discovery curve of rMLST alleles over rST profiles varied by species, as shown in Figure 1. For *M. tuberculosis* and *S. aureus*, the relationship between discovered alleles and rST profiles was effectively linear, in contrast to *C. jejuni* and *N. meningitidis*, where fewer alleles are discovered as the number of rST profiles increases, indicating a higher number of profiles using novel combinations of alleles previously seen in other profiles.

Table 2. Profiles and total alleles in rMLST sequences

	Ribosomal isolates	**rST profiles**	**Total rMLST alleles**	**Allele-to-profile ratio**
C. jejuni	15,654	3,028	6,226	2.056
M. tuberculosis	12,501	1,118	2,006	1.794
N. meningitidis	11,840	3,267	3,436	1.052
S. aureus	33,150	4,524	6,514	1.440

Figures 1a and 1b. Discovery curves of total alleles in rMLST loci found in four species, by isolate count and rST count.

Isolates for each species were separated into 1000 random sets of equal size and in random order. The number of unique alleles in total over 53 loci and the total number of rST profiles were assessed for each species, for the first set, for the first two sets, then for first three sets, and so forth cumulatively until all 1000 random sets were included. Each step of subdivision was plotted to obtain the discovery curve, both by the number of isolates in the sets collectively, shown in 1a, and by the number of unique rST profiles found in the sets, shown in 1b.

Figure 1a.

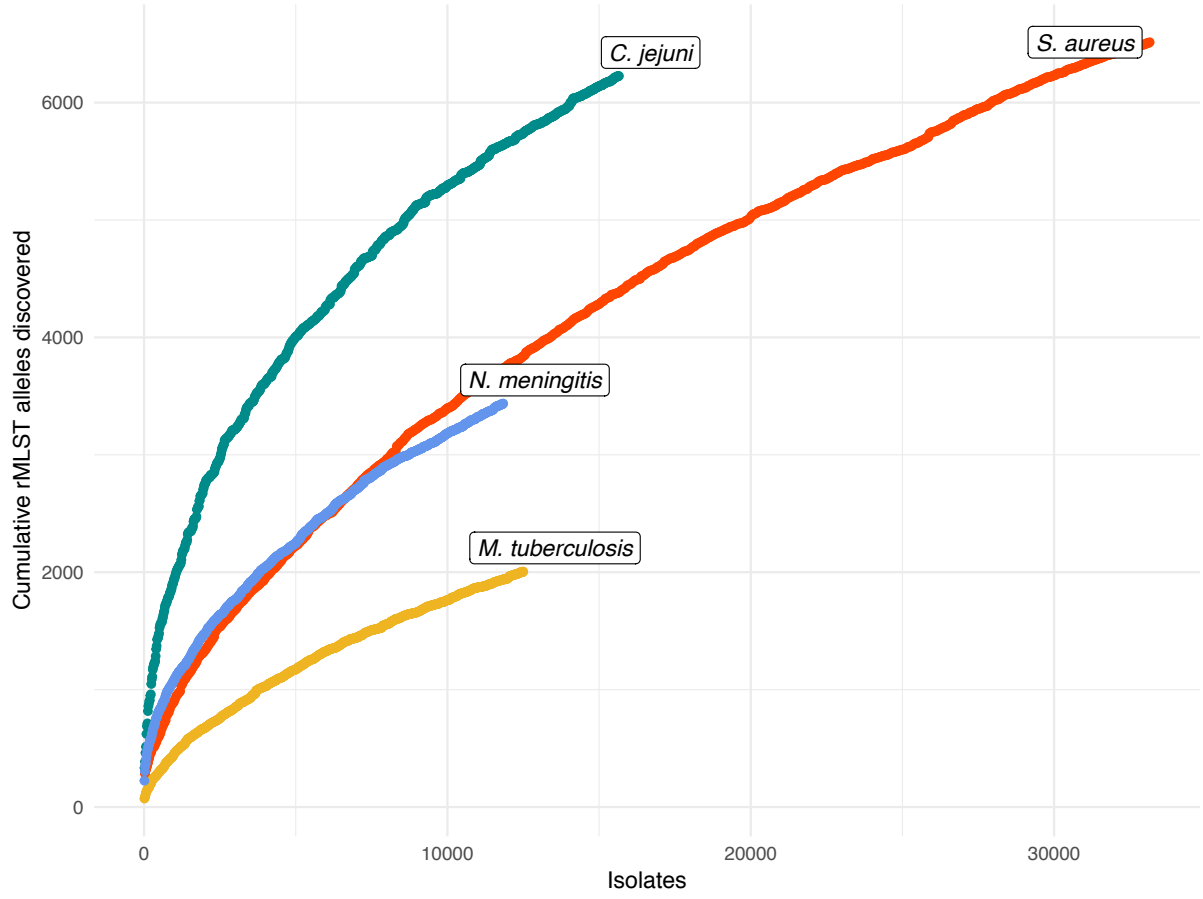
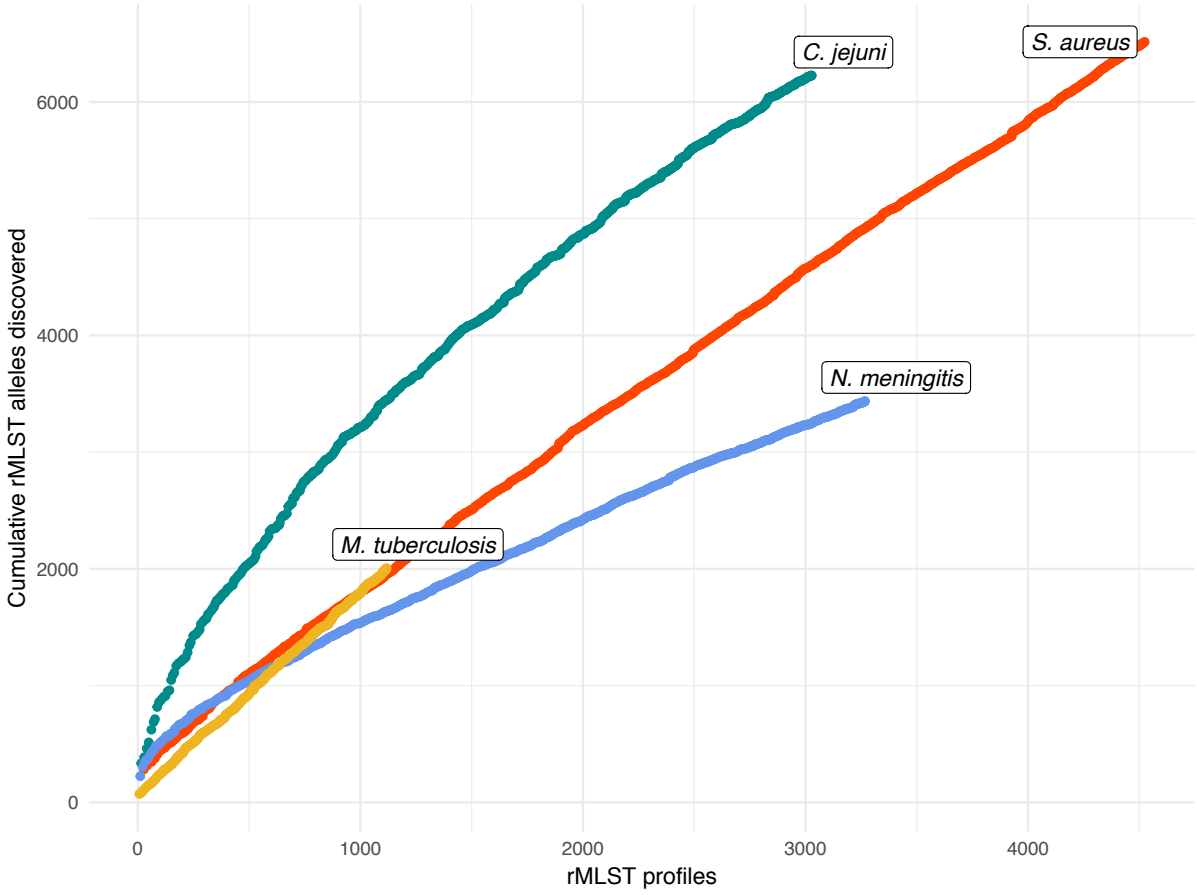


Figure 1b.


C. jejuni
S. aureus
N. meningitis
M. tuberculosis
Cumulative rMLST alleles discovered
rMLST profiles
0
2000
4000
6000
0
1000
2000
3000
4000

The genetic diversity within the rMLST alleles of each species was measured, as shown in Table 4 and Figure 2. *M. tuberculosis* had the lowest diversity, with an average of 0.092 alleles per base pair, while *C. jejuni* had the maximum, at 0.288 alleles per base pair. Each species had genetic diversity means significantly different from others, even after including locus as a blocking factor, with the exception of the comparison between *C. jejuni* and *S. aureus* ($p = 0.850$).

This measure of genetic diversity is influenced by the quantity and genetic distance between samples. Both of these factors vary in the four species for which isolates were collected in this study. For example, the "Neisseria PubMLST database" includes many carriage isolates, while the "C. jejuni/coli database" focuses largely on human disease isolates, with relatively few environmental isolates or isolates from other hosts.

Loci also varied in their genetic diversity, with a minimum value of 0.108 alleles per base pair for *rpsN* and a maximum for 0.307 for *rplI*, with the mean of the 53 loci at 0.208 with a standard deviation of 0.116 when including paralogous loci, or a mean of 0.212 and a standard deviation of 0.113 when paralogous loci are excluded. The differences between loci were not significant however, with no relationship found in the ranks of loci as ordered by their allelic diversity ($p = 0.623$), or the means from the genes corresponding to proteins in the large and small subunits of the ribosome ($p = 0.248$). Differences in diversity between loci do not depend on the diversity of isolates as all loci are present in most isolates, with an average of 158.76 allele designations missing per locus over 73,145 isolates, or 0.002%.

N. meningitidis genomes and S. aureus genomes have two pairs of paralogous loci, while *M. tuberculosis* genomes have four pairs and *C. jejuni* loci have none. For *N. meningitidis*, these paralogous loci were among the most diverse in the genome, but in *S. aureus* and *M. tuberculosis*, paralogous loci ranked among the least diverse.

Table 3. Allelic diversity in rMLST genes, measured in alleles per nucleotide.

Species	Average	Standard deviation
C. jejuni	0.288	0.079
M. tuberculosis	0.092	0.019
N. meningitidis	0.155	0.083
S. aureus	0.303	0.081
Overall	0.212	0.113

Figure 2. Allelic diversity in the rMLST loci over four species.

Boxplots show the mean (central black line), and first and third quantiles (as the edges of the box) values of allelic diversity, in log-transformed alleles per base pair length, for four species. Box whiskers extend to the minimum or maximum values, up to 1.5 times the inter-quantile length. Outliers beyond this are shown as single black points. All pairwise comparisons found significantly different means across species ($p < 0.001$), except for that between *C. jejuni* and *S. aureus* ($p = 0.850$) .

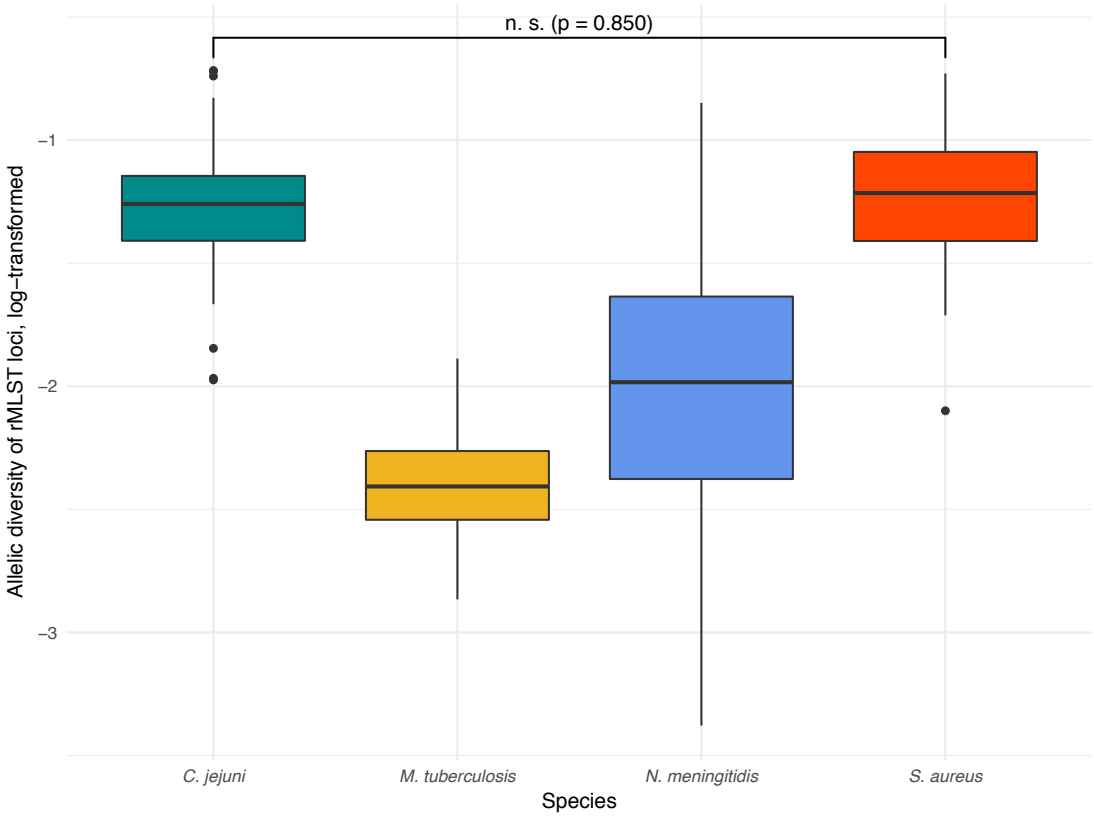
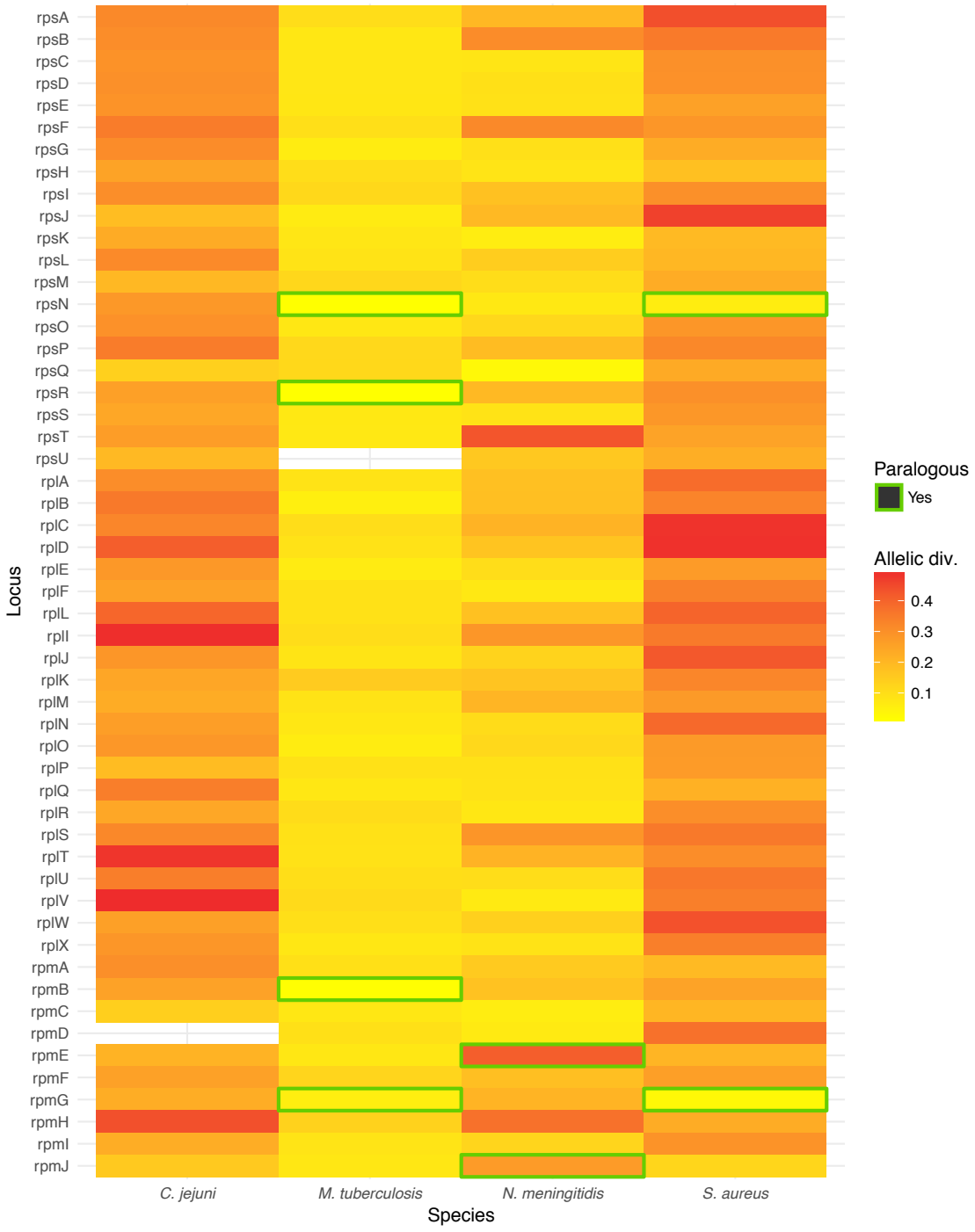


Figure 3. Allelic diversity in the 53 rMLST loci over four species.

A heatmap shows allelic diversity, measured as number of unique alleles divided by average allele length in base pairs. All 53 loci in the scheme are included, with absent loci marked by a white cell and paralogous loci identified with a green boundary. The heat map colours range from yellow (minimum) to red (maximum), with decreasing brightness. The minimum value was 0.015 and the maximum was 0.488.



Unlike genetic diversity, the allele ratio is not dependent on the number of isolate samples except for the former's effect on the power of the analysis. Since the allele ratio is a ratio between two values within the same dataset, the number of unique sequences after and before translation, it is inherently a paired test.

Allele ratio values varied between species, as shown in Table 5 and Figure 4. *C. jejuni* had the lowest ratio, at 0.376 unique amino acid sequences per unique nucleotide sequence, while *M. tuberculosis* had the highest, at a ratio of 0.635. The mean across all species was a ratio of 0.487, with a standard deviation of 0.150. In all pairs of species, the means of their allele ratios over the 53 rMLST loci were significantly different from one another ($p < 0.001$), with the exception of *N. meningitidis* and *C. jejuni* ($p = 0.202$).

As with allelic diversity, no significant relationship was found in the rank of loci by allele ratio over the four species ($p = 0.805$). However, the allele ratio in the small and large subunits of the ribosome had means that differed significantly from one another ($p = 0.045$), with means of 0.467 and 0.513 respectively.

Table 4. Selection measures for rMLST loci in four species.

Species	**Mean**	**Standard deviation**
C. jejuni	0.376	0.121
M. tuberculosis	0.635	0.082
N. meningitidis	0.413	0.138
S. aureus	0.536	0.089
Overall	0.487	0.150

Figure 4. Allele ratios in the rMLST loci over four species.

Boxplots show the mean (central black line), and first and third quantiles (as the edges of the box) allele ratio values for four species. Box whiskers extend to the minimum or maximum values, up to 1.5 times the inter-quantile length. Outliers beyond this are shown as single black points. All pairwise comparisons found significantly different means across species ($p < 0.001$), except for that between *C. jejuni* and *N. meningitidis* ($p = 0.202$) .

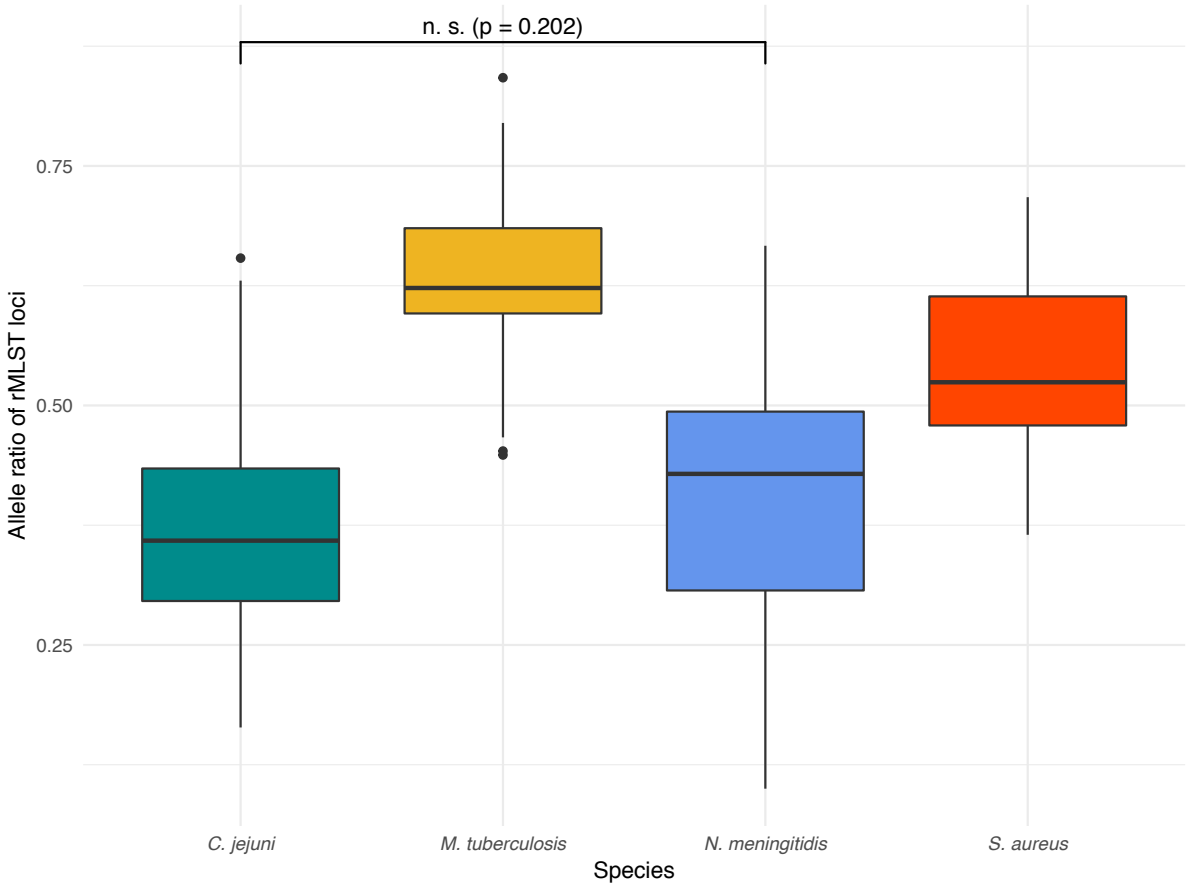
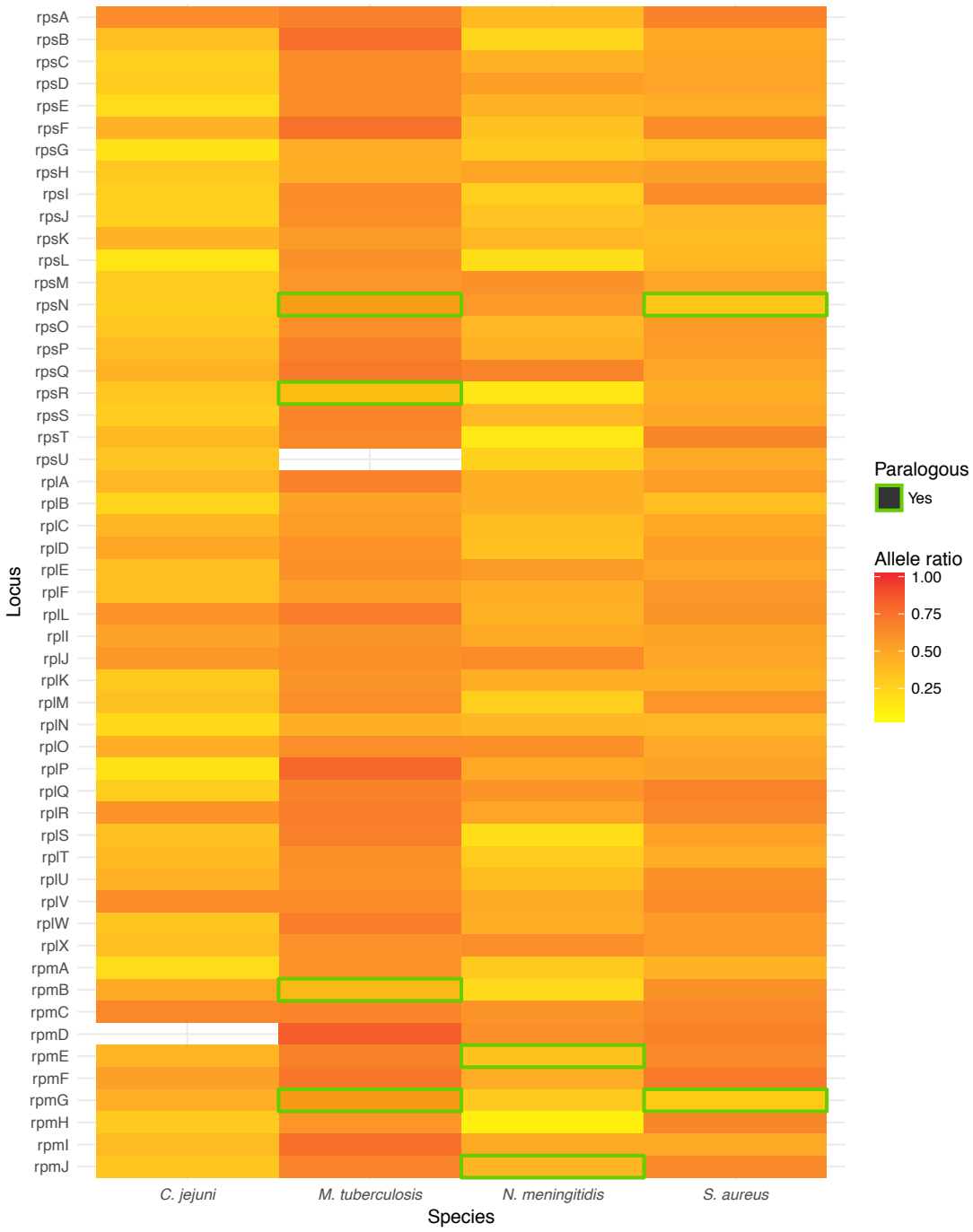


Figure 5. A heatmap shows allelic diversity, measured as ratio of unique alleles sequences after and before translation.

All 53 loci in the scheme are included, with absent loci marked by a white cell and paralogous loci identified with a green boundary. The heat map colouring ranges from yellow (minimum) to red (maximum), with decreasing brightness. The minimum value was 0.010 and the maximum was 1.00.



Loci in the core genome of each species were assessed for their average allele ratio values, as shown in Table 5 and Figure 6. The number of core loci varied by species, from 1,375 loci in *C. jejuni* to 3,172 loci in *M. tuberculosis*. *M. tuberculosis* and *S. aureus* both showed relatively high ratios, with means of 0.647 and 0.644 unique amino acid sequences per unique nucleotide sequence respectively. The means of these two species were not significantly different from one another ($p = 0.694$). In contrast, both species were significantly different from the remaining two species, *C. jejuni* and *N. meningitidis*, which both had lower ratios, at means of 0.554 and 0.544 respectively. *C. jejuni* and *N. meningitis* had similar ratios, with no significant difference in means ($p = 0.067$).

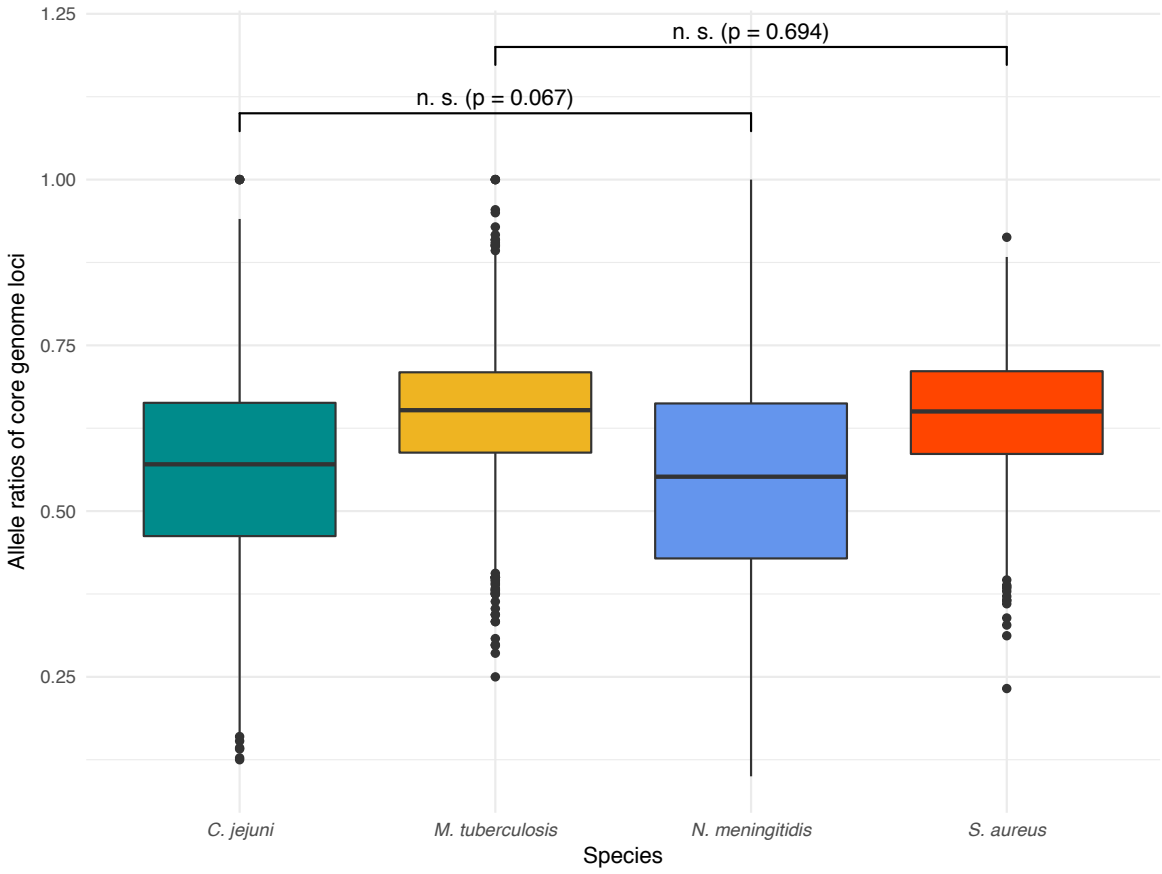
These relationships between species mirrored the relationships at the rMLST loci, where *M. tuberculosis* displayed the highest ratios, followed by *S. aureus* and finally *C. jejuni* and *N. meningitidis*, who had similar low ratios. Allele ratios were all generally higher for core genome loci than for rMLST loci, with all species having higher values in the core genome.

Table 5. Selection measures for core genome loci in four species.

Species	**Average allele ratio**	**Standard deviation**
C. jejuni	0.554	0.128
M. tuberculosis	0.647	0.145
N. meningitidis	0.544	0.096
S. aureus	0.644	0.163

Figure 6. Allele ratios in the core genome loci of four species.

Boxplots show the mean (central black line), and 1st and 3rd quantile (as the edges of the box) allele ratio values for four species. Box whiskers extend to the minimum or maximum values, up to 1.5 times the inter-quantile length. Outliers beyond this are shown as single black points. All pairwise comparisons found significantly different means across species ($p < 0.001$), except for two, that between *C. jejuni* and *N. meningitidis* ($p = 0.067$) and between *M. tuberculosis* and *S. aureus* ($p = 0.694$).



Discussion

The four species in this analysis, *C. jejuni*, *M. tuberculosis*, *N. meningitidis* and *S. aureus*, are all widespread human pathogens. Despite this key similarity, they are species with varied host ranges, disease severity and lifestyles.

Campylobacter jejuni and a closely related species with which it can sometimes recombine, *Campylobacter coli,* are ubiquitous in nature, in the environment and in animals, where they are often carried asymptomatically. In contrast, when infecting humans they cause diarrhoeal disease. They frequently cause of disease in humans, infecting around 1% of the population of Western Europe each year (Humphrey et al. 2007). A majority of human cases are caused by infection from poultry meat, though other domestic animals and wild birds can also be sources (Cody et al. 2015). This wide host range and large reservoirs, combined with the high levels of HGT display in both the core and accessory genomes of *Campylobacter* cause it to a highly adaptable species (Sheppard et al. 2013).

Mycobacterium tuberculosis contrasts with *C. jejuni* on many of these features. It has no reservoir, as it is an obligate human pathogen, and it does not undergo any HGT, leading it to be a species with low genetic diversity and a highly structured population with distinct lineages and sub-lineages (Niemann et al. 2016). In fact, WGS is the only molecular technique with sufficient resolution to accurately classify strains, and additionally to uncover the genetic determinants of virulence (Coscolla and Gagneux 2014). This technology has led to the national-level genotyping programs that use WGS to identify transmission events, detect drug resistance mutations and reveal phylogenetic structures (Nikolayevskyy et al. 2016). Large-scale efforts such as these are necessary to combat *M. tuberculosis*, as it has the highest disease burden of any bacteria. It causes over a millions deaths each year, a number that belies its even more massive prevalence, as around two billion people around the world carry *M. tuberculosis* in its dormant state. An *M. tuberculosis* infection may be carried

asymptomatically for decades before it becomes activated and causes disease, typically in the respiratory system (Gagneux 2012).

While *M. tuberculosis* transmission and virulence are closely linked, *Neisseria meningitidis* may best be classed as an "accidental pathogen". *N. meningitidis* can cause meningococcal disease, but it is found in the human nasopharynx as a harmless commensal organism all around the world. Invading the human body leads it to an evolutionary dead-end from which it is unlikely to transmit to another host (Maiden and Harrison 2016). Not all strains of *N. meningitis* are equally likely to be associated with disease, but genetic determinants of the invasive phenotype remain unclear (Bratcher et al. 2012). The frequent rates of HGT, mostly intra-species though some inter-species HGT has been observed, for example between *N. meningitidis* and *N. lactamica* (Mulhall et al. 2016), have complicated research into the determinants of pathogenicity.

For *Staphylococcus aureus*, the genetic factors contributing to pathogenicity are much clearer. The species is perhaps most famous as the bacteria in "MRSA," the acronym given to its methicillin-resistant strains, hence "MR". *S. aureus* can cause a wide variety of infections in humans, but it is the leading agent of bacteremia and endocarditis, as well as a common cause of skin and soft tissues infection. Like *N. meningitidis*, it can live as a commensal organism in the human microbiome and can be found in around 30% of the world's population as such a commensal (Tong et al. 2015). Virulence factors and drug resistance genes can be found in mobile elements of the *S. aureus* genome, including phages, transposons, genomic islands and plasmids. For these components of the accessory genome, HGT occurs frequently between *S. aureus* species. Meanwhile, the core genome of *S. aureus* is subject to HGT much more rarely. This combination causes the species' phylogeny to clearly structured into clonal complexes determined by vertical descent despite a highly variable accessory genome (Planet et al. 2017).

The difference in evolutionary mechanisms in use by these species affects the structure of their populations, which can be observed by differences in their patterns of genetic diversity. Genetic diversity can be increased in a bacterial population by two means: mutation and HGT. The rates of both methods vary by species and even within different parts of a genome, as is the case with the core and accessory genes in *S. aureus*. The interplay between these mechanisms of diversity generation affects the structure of diversity with bacterial populations.

In ecological studies, the total number of extant species can be extrapolated from the pattern of discovery of new species. Discovery efforts over time can be analysed at once to form a species accumulation curve, which rises sharply at first and then approaches an asymptote that represents the total number of species in the taxonomic group in question (Bebber et al. 2007). The discovery curves plotted for the total rMLST alleles profiles based on their records in the PubMLST database indicate that these collections remain too incomplete to accurately predict an asymptote, and therefore predict the extant total of alleles. This may be due to insufficient sampling, or it may point to fundamental differences in the evolutionary dynamics of macrobiological species and microbiological alleles. For one, it is possible that new alleles appear at much higher rates than new species, so that the number of alleles available for discovery in a population is not fixed during the surveying period. If this is the case, as isolates are collected over years or decades, they are drawn from populations whose extant member differ in presence or at least in distribution. This is a possibility, as *S. aureus*, for example, is known to undergo "strain replacement", where one strain is most prevalent at any given moment, and can be replaced by another over a period of years or decades (Albrecht et al. 2011).

The relationship between the number of rST profiles and total alleles discovered pointed to another feature of evolution: recombination in core genome alleles. The relationship between total number of rST profiles and total number of rMLST alleles was linear for

M. tuberculosis and *S. aureus*, indicating that the evolution of their ribosomal genes is clonal, defined largely by vertical descent. For *C. jejuni* and *N. meningitidis*, rST profiles displayed greater numbers of repeated alleles appearing in novel combinations, pointing to HGT that affects even the most conserved loci of the core genome. This is consistent with the existing literature in these species, where *M. tuberculosis* is believed to be fully clonal, *S. aureus* undergoes little HGT in its core genome, and *C. jejuni* and *N. meningitidis* are both subject to frequent HGT, including inter-species recombination.

The introduction of novel genotypes into a genome through HGT comes with costs and benefits. HGT can remove the deleterious effects caused by mutations, by introducing un-mutated genotypes transferred from other bacteria, or it can provide entirely novel alleles or genes with beneficial fitness effects. However, as genomes are collections of genes that have evolved together for many generations, adapting collectively to their specific niche. There is therefore a possibility that a novel introduction will cause a disturbance with deleterious effects.

The rate of synonymous to non-synonymous mutations can reveal something about the selection forces acting upon a population. Synonymous mutations, which do not cause any change to the chain of amino acids that forms the backbone of translated proteins, rarely have an effect on the biology of an organism. Therefore, in most cases, synonymous mutations are virtually neutral in terms of fitness. Non-synonymous mutations, however, do affect the protein's amino acid chain, and can have an effect on protein structure. A random change to protein structure will generally be detrimental to its function, causing non-synonymous mutations to be largely deleterious. Therefore the rate of synonymous mutations can act as a relatively neutral benchmark, against which the rate of non-synonymous and therefore relatively impactful mutations can be tracked. The removal of non-synonymous mutations from a gene indicates that it is functionally conserved, while an over-abundance of non-synonymous mutations can indicate a gene undergoing positive

selection (Kryazhimskiy and Plotkin 2008). In terms of bacteria, functional conservation is often observed in "housekeeping" genes with critical functions like DNA replication and protein synthesis, while positive selection can be observed in pathogens when antigens that interact with host immune systems constantly evolve to evade immune defenses.

The rate of non-synonymous to synonymous mutations was approximated in this analysis as the allele ratio, the ratio between unique alleles sequences in amino acid and in nucleotide format. A higher rate of non-synonymous mutations would increase this ratio, as it would increase the number of allele sequences with unique amino acid sequences. The average allele ratio differed significantly between the species in this analysis, in both their ribosomal genes and more widely in their species-specific core loci. The species with the lowest rates of recombination in their core genes, as predicted by the linear relationship between their total alleles and number of rST profiles, namely *M. tuberculosis* and *S. aureus*, also displayed the highest allele ratios in both rMLST and core genome loci. *M. tuberculosis* is known to have elevated rate of synonymous mutations, due to its small effective population sizes, a consequence of its lack of HGT (Hershberg et al. 2008). *S. aureus* has likewise been shown to have largely vertical descent in its core genome, with very limited rates of HGT (Planet et al. 2017). *C. jejuni* and *N. meningitidis*, in contrast, are highly recombinant.

The relationship between genetic diversity and the degree of HGT as known in the literature for each species was not as straightforward. For example, *N. meningitis* showed a low level of diversity despite its reported high levels of recombination. Its sequence diversity over the rMLST loci is significantly lower than that of *C. jejuni*, despite a similar number of isolates and similar number of rSTs. This could be a reflection of a stricter species definition for *N. meningitidis* than for *C. jejuni*, as the criteria that determine the boundaries of bacterial species are not applied universally (Lan and Reeves 2001; Bapteste and Boucher 2009), and can be especially loosely defined for pathogens (Stackebrandt et al. 2002). Differences in observed diversity could also be due to sampling differences. *S. aureus* and *C. jejuni* showed a

similar level of genetic diversity over the rMLST loci, despite nearly three times as many isolates of the former being present in the database. This could indicate a higher diversity in the *C. jejuni* species as compared to *S. aureus*. Alternatively, it could reflect the poor representation of *S. aureus* species diversity in the database due to a sampling bias toward pathogenic strains. Strains with higher medical or epidemiological relevance are more likely to be sampled and sequenced, and this is known to be true in the case of publicly available *S. aureus* sequences (Planet et al. 2017). In this scenario, the genetic diversity displayed by the *S. aureus* isolates in the PubMLST database may be representative of a subset of the *S. aureus* population, while species for which dedicated efforts to increase the sampling of non-disease-causing isolates, as is the case for *Neisseria* and *Campylobacter*, may have collections which are more truly representative of their species.

Nevertheless, the differences observed in this dataset reflect known characteristics of these species. The genome of *M. tuberculosis* has low sequence diversity, to the level where discrimination of strains was difficult to achieve before WGS and not always possible even with WGS (Nikolayevskyy et al. 2016). *S. aureus* in contrast, despite relying on mutation rather than HGT to increase the diversity of its core genome, is known to have moderately high rates of mutation (Hanage et al. 2006).

The allele-based methods explored here present a rapid and scalable method by which the diversity and selection signals in large data sets, containing tens of thousands of isolates can be investigated. As WGS continues to be used for clinical or epidemiological studies, whole genome sequences for more bacteria, of more diverse species will be collected. These can be investigated in the manner shown here, despite their original use, to collectively determine characteristics of the population from which they were drawn.

This analysis additionally supports the statement that the ribosomal protein genes are under similar evolutionary forces in multiple species. The loci in the rMLST scheme did not appear to differ in their rates of accumulated diversity in similar ways in distinct species, and only a slightly effect was found in the strength of selection acting on the two subunits, where the small subunit was found to be under slightly stronger purifying selection.

Supplemental files

S1 Table containing allele designations for the rMLST loci over 73,145 isolates.

S2 Table containing GUAVA results for the rMLST loci at each of four species, as Excel file with multiple sheets.

S3 Table containing GUAVA results for the rMLST loci at each of four species, as Excel file with multiple sheets.

S4 Table containing GUAVA results for the cgMLST loci at each of *C. jejuni*.

S5 Table containing GUAVA results for the cgMLST loci at each of *M. tuberculosis*.

S6 Table containing GUAVA results for the cgMLST loci at each of *N. meningitidis*.

S7 Table containing GUAVA results for the cgMLST loci at each of *S. aureus*.

S8 Table containing accumulation curve results for isolates, rSTs and rMLST alleles over four species.

Chapter IV ***References***

Albrecht N, Jatzwauk L, Slickers P, et al (2011) Clonal Replacement of Epidemic Methicillin-Resistant Staphylococcus aureus Strains in a German University Hospital over a Period of Eleven Years. PloS One. doi: 10.1371/journal.pone.0028189

Bapteste E, Boucher Y (2009) Epistemological impacts of horizontal gene transfer on classification in microbiology. Methods Mol Biol 532:55–72. doi: 10.1007/978-1-60327-853-9_4

Bebber DP, Marriott FHC, Gaston KJ, et al (2007) Predicting unknown species numbers using discovery curves. Proceedings Biological sciences / The Royal Society 274:1651–1658. doi: 10.1098/rspb.2007.0464

Bratcher HB, Bennett JS, Maiden MCJ (2012) Evolutionary and genomic insights into meningococcal biology. Future Microbiol 7:873–885. doi: 10.2217/fmb.12.62

Bratcher HB, Corton C, Jolley KA, et al (2014) A gene-by-gene population genomics platform: de novo assembly, annotation and genealogical analysis of 108 representative Neisseria meningitidis genomes. BMC Genomics 15:1138. doi: 10.1186/1471-2164-15-1138

Cody AJ, Bray JE, Jolley KA, et al (2017) Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of Campylobacter jejuni and C. coli Human Disease Isolates. J Clin Microbiol 55:2086–2097. doi: 10.1128/JCM.00080-17

Cody AJ, McCarthy ND, Bray JE, et al (2015) Wild bird-associated Campylobacter jejuni isolates are a consistent source of human disease, in Oxfordshire, United Kingdom. Environ Microbiol Rep 7:782–788. doi: 10.1111/1758-2229.12314

Cody AJ, McCarthy ND, Jansen van Rensburg M, et al (2013) Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol 51:2526–2534. doi: 10.1128/JCM.00066-13

Coscolla M, Gagneux S (2014) Consequences of genomic diversity in Mycobacterium tuberculosis. Semin Immunol 26:431–444. doi: 10.1016/j.smim.2014.09.012

Gagneux S (2012) Host-pathogen coevolution in human tuberculosis. Phil Trans R Soc B 367:850–859. doi: 10.1098/rstb.2011.0316

Hanage WP, Fraser C, Spratt BG (2006) The impact of homologous recombination on the generation of diversity in bacteria. J Theor Biol 239:210–219. doi: 10.1016/j.jtbi.2005.08.035

Hershberg R, Lipatov M, Small PM, et al (2008) High Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human Demography. PLoS Biol 6:e311. doi: 10.1371/journal.pbio.0060311

Humphrey T, O'Brien S, Madsen M (2007) Campylobacters as zoonotic pathogens: a food production perspective. Int J Food Microbiol 117:237–257. doi: 10.1016/j.ijfoodmicro.2007.01.006

Jansen van Rensburg MJ, Swift C, Cody AJ, et al (2016) Exploiting Bacterial Whole-Genome Sequencing Data for Evaluation of Diagnostic Assays: Campylobacter Species Identification as a Case Study. J Clin Microbiol 54:2882–2890. doi: 10.1128/JCM.01522-16

Jolley KA, Bennett JS, Maiden MCJ, et al (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiol 158:1005–1015. doi: 10.1099/mic.0.055459-0

Jolley KA, Maiden MCJ (2013) Automated extraction of typing information for bacterial pathogens from whole genome sequence data: Neisseria meningitidis as an exemplar. Euro Surveill 18:20379–.

Köser CU, Niemann S, Summers DK, Archer JAC (2012) Overview of errors in the reference sequence and annotation of Mycobacterium tuberculosis H37Rv, and variation amongst its isolates. Infect Genet Evol 12:807–810. doi: 10.1016/j.meegid.2011.06.011

Kryazhimskiy S, Plotkin JB (2008) The Population Genetics of dN/dS. PLoS Genetics. doi: 10.1371/journal.pgen.1000304

Lan R, Reeves PR (2001) When does a clone deserve a name? A perspective on bacterial species based on population genetics. Trends Microbiol 9:419–424.

Maiden M, Bygraves JA, Feil E, et al (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145. doi: 10.1073/pnas.95.6.3140

Maiden MCJ, Harrison OB (2016) The population and functional genomics of the Neisseria revealed with gene-by-gene approaches. J Clin Microbiol 54:JCM.00301–16–1955. doi: 10.1128/JCM.00301-16

Maiden MCJ, Jolley KA (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595. doi: 10.1186/1471-2105-11-595

Maiden MCJ, van Rensburg MJJ, Bray JE, et al (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:728–736. doi: 10.1038/nrmicro3093

Meric G, Miragaia M, de Been M, et al (2015) Ecological Overlap and Horizontal Gene Transfer in Staphylococcus aureus and Staphylococcus epidermidis. Genome Biol Evol 7:1313–1328. doi: 10.1093/gbe/evv066

Mulhall RM, Brehony C, O'Connor L, et al (2016) Resolution of a Protracted Serogroup B Meningococcal Outbreak with Whole-Genome Sequencing Shows Interspecies Genetic Transfer. J Clin Microbiol 54:2891–2899. doi: 10.1128/JCM.00881-16

Niemann S, Merker M, Kohl T, Supply P (2016) Impact of Genetic Diversity on the Biology of Mycobacterium tuberculosis Complex Strains. doi: 10.1128/microbiolspec.TBTB2-0022-2016

Nikolayevskyy V, Kranzer K, Niemann S, Drobniewski F (2016) Whole genome sequencing of Mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: A systematic review. Tuberculosis (Edinb) 98:77–85. doi: 10.1016/j.tube.2016.02.009

Ohta T (1992) The Nearly Neutral Theory of Molecular Evolution. Annual Review of Ecology and Systematics 23:263–286. doi: 10.2307/2097289

Ohta T (2011) Near-neutrality, robustness, and epigenetics. Genome Biol Evol 3:1034–1038. doi: 10.1093/gbe/evr012

Planet PJ, Narechania A, Chen L, et al (2017) Architecture of a Species: Phylogenomics of Staphylococcus aureus. Trends Microbiol 25:153–166. doi: 10.1016/j.tim.2016.09.009

Salipante SJ, SenGupta DJ, Cummings LA, et al (2015) Application of Whole-Genome Sequencing for Bacterial Strain Typing in Molecular Epidemiology. J Clin Microbiol 53:1072–1079. doi: 10.1128/JCM.03385-14

Sheppard SK, Didelot X, Jolley KA, et al (2013) Progressive genome-wide introgression in agricultural Campylobacter coli. Mol Ecol 22:1051–1064. doi: 10.1111/mec.12162

Sintchenko V, Iredell JR, Gilbert GL (2007) Pathogen profiling for disease management and surveillance. Nat Rev Microbiol 5:464–470. doi: 10.1038/nrmicro1656

Stackebrandt E, Frederiksen W, Garrity GM, et al (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043–1047. doi: 10.1099/00207713-52-3-1043

Yahara K, Didelot X, Jolley KA, et al (2016) The Landscape of Realized Homologous Recombination in Pathogenic Bacteria. Mol Biol Evol 33:456–471. doi: 10.1093/molbev/msv237

Tong SYC, Davis JS, Eichenberger E, et al (2015) Staphylococcus aureus Infections: Epidemiology, Pathophysiology, Clinical Manifestations, and Management. Clin Microbiol Rev 28:603–661. doi: 10.1128/CMR.00134-14

General Discussion

Table of Contents

Discussion **1**

General Conclusion **6**

General Discussion **References** **7**

Discussion

Whole genome sequencing has caused a seismic shift in the quantities of data that exist in the world of bacteriology. For example, the first whole genome sequence of *Mycobacterium tuberculosis* was published in 1998 (Cole et al. 1998). The three dozen authors credited for that work belie the incredible effort required at the time to produce and annotate just one whole genome sequence. Not even twenty years later, this work, produced largely by a single student, includes the analysis of over seven thousand *M. tuberculosis* isolates.

However, the mere existence of data is not sufficient to advance the edge of human knowledge. It is a first step that must be followed by interpretation, for which data analysis is a necessary. When the first DNA fragments were sequenced, samples that required hours upon hours of laboratory time could nevertheless fit in a single page. Strings of A, C, G and T could be aligned by hand, in the cases where a sequence against which they could be aligned existed at all (McLafferty et al. 1988). In contrast, Chapter III in this work contains a dataset of 24,403,680 allele designations. Manual analysis of data in this scale is unthinkable. Using a rather conservative estimate wherein manual assignment of each allele designation requires

10 seconds to be completed, it would take 32 years of continuous 40-hour workweeks for one researcher to create a dataset of that size. It bears noting that this dataset is not the only one presented in this work, and neither is it the largest. Automation is simply no longer an option in the study of bacterial genomes; it is a necessity.

That is not to say that the human element of curation in these datasets can be discarded entirely. Bacteria are living things, and nothing in biology is perfectly predictable. The supposedly universal typing scheme based on ribosomal genes that is the focus of Chapter IV presents an example of this. The ribosome is a critical part of the cellular machinery and possibly one of the most conserved structures in all of life (Lecompte et al. 2002). Its 53 proteins, however, are not all present in the species reviewed. In fact, only 45 of these 53 hyper-conserved genes are present in the same copy number in all four species.

A fully automated method may have made a number of critical mistakes when encountering a dataset such as this. In the case of the two copies of *rpmG* in *M. tuberculosis*, for example, it may have simply stopped searching after finding the first copy, or it may have determined that all isolates contained sequences from two strains and discarded them as a form of contamination control. It is only with a combination of expert human input, charged with the power of automation, that datasets of this size can be effectively managed by tailoring them to the quirks of each individual bacterial species. Only then will the most accurate and complete results possible be created.

Accurate annotation of genomes requires the expertise that combines knowledge about an organism with the knowledge of effective data analysis. The potential of this combination is not new to the field of bacteriology. Recall the famous case of the 1849 cholera outbreak in London. A map of deaths by cholera only led to the accurate identification of the neighbourhood water pump as the source of infection because Dr. John Snow knew something about the lifestyle of *Vibrio cholerae,* specifically its lifestyle as a water-borne pathogenic bacterium (Colwell 1996). Likewise, today the annotation of whole genome

sequences is an exercise that requires the understanding of the nature of genetic variation in bacteria in order to achieve its full potential.

However, the time of such experts is finite. Techniques must be developed with this limitation in mind, and strive to make the most use of their efforts. The gene-by-gene method and its basis on the annotation of genomes in a permanent and portable way presents one possible solution to this. Annotations can be cumbersome to create, but they do not need to be changed when more genomes are added to a collection, or as new genes are defined. This permanence avoids the need for repetitive work.

This need is all the more pressing in light of the deluge of data under which we find ourselves now. As further improvements are made to the technologies that underpin WGS, speed and cost will continue to improve. The quality of the data produced will likewise increase as read lengths and read accuracy climb. As the convenience of using WGS rises, with more portable, cheaper machines that require less preparatory time, more sequences will be created from a greater variety of contexts. These improvements to the supply of WGS are matched by increase in its demand (Goodwin et al. 2016). WGS is already used for tracking transmission, identifying outbreaks, predicting antibiotic resistance profiles and tracing the evolutionary path of an organism (Nikolayevskyy et al. 2016).

The hope for this thesis is that is provides a method by which existing data and existing models are combined in a novel way to increase our knowledge about the biological features of bacterial species.

Allele-based methods were used extensively in this work, but they are not the only methods available. Alignment-based methods may be more computationally intensive, but they do provide the highest resolution (Didelot and Falush 2007). While the maximum possible level of discrimination between isolates is not required in most contexts, for fine-scale analyses of transmission, it can be necessary. Allele-based and alignment-based

methods are not in competition but can — and if fact should — continue to be developed and used alongside one another.

In fact, both allele-based and alignment-based methods are used in the analyses of the genome of *M. tuberculosis* presented in this work. As the most extensively studied fully clonal pathogen, this species represents a key model for the genomics of clonal bacteria, of pathogenic bacteria, and of bacteria as a whole. I have presented here a novel method, developed for the express purpose of investigating the variation present in this low-diversity species in the maximum possible number of isolates at once. This makes it possible to extract the most knowledge from data that has already been created, improving our understanding of this widespread human pathogen at a low cost. In many cases, the results from these analyses, despite the use of novel methods, confirmed the existing literature. For example, they identified several antibiotic resistance genes by their high levels of homoplasy between the two lineages most often associated with the rise of drug resistance. A multi-species analysis reinforced the standing of *M. tuberculosis* as a species with extremely low genetic diversity and a clonal population structure. This low diversity can be deceptive, however. Diversity may be low on average, but it is highly variable between individual genes in the genome. The most diverse genes had far higher numbers of alleles and variable sites than comparable genes with similar sizes. Similarly, high variability was found in the rate of non-synonymous to synonymous mutations, despite the genome evolving through solely through genome-wide sweeps.

The analyses presented here point to clear avenues of further work. For one, Chapter III identifies dozens of genes within *M. tuberculosis* that show distinct patterns of evolution apart from the majority of the genome, but have no known function. The same patterns are observed in many of the genes known to be involved in virulence or drug resistance, lending credence to the idea that these so far unexplored genes may have important functions. Chapter IV has explored the dynamics of four well-known bacterial pathogens, but can

readily be expanded to over a multitude of species, especially as more species come to be represented by large numbers of isolates. Chapter II presents a piece of software designed specifically to be quick to set up and quick to perform its analyses. The inclusion of a visualisation package also boosts its utility as a user-friendly and lightweight application that can minimise the time between the acquisition of data and the interpretation of results. It is my hope that it comes to be used widely.

There is also a pressing need for more complex forms of data. Only genetic sequences were used in this work, in the absence of any "meta data" that described the characteristics of the host or environment in which the sample was collected, or even any phenotypic traits of the bacteria themselves. Datasets such as these are sufficient for some analyses, as the various mechanisms acting on genetic sequences leave distinct marks on the genome. However, a more complete data set, that combines the ultimate genetic resolution of WGS and knowledge about physiology of the organism it describes, would have considerably more power.

General Conclusion

Whole genome sequencing is already in widespread use in the study of bacteria, and the frequency of this use is set to continue increasing at a rapid face. The large data sets that result from these growing collections of whole genome sequences require analytical methods tailored to their scale. These methods must allow human curation and automation to complement one another, so as to make the most of the benefits of human expertise and of computational power.

As the collections currently stored in PubMLST demonstrate, isolate collections numbering in the hundreds of thousands of whole genome sequences already exist. In many cases, these have already been curated by human experts, or exist in customized databases that make the most of existing knowledge about the species or genus contained within it. Methods that can extract new findings from these collections are therefore needed now.

I have developed a method that fits these criteria and is now available as software that anyone can use, as detailed in Chapter II. It emphasis on rapid analysis and the inclusion of a visualisation package make it especially well suitable for wide application.

Chapter III contains an extensive application of this method to a large collection of *Mycobacterium tuberculosis* genomes. It confirmed some known effects, and pointed to dozens of genes that have so far remain unexplored as potential sites of interest in the genome. The method was similarly applied to three further species with varied lifestyles in Chapter IV, uncovering the marks left in their genomes by the effects of mutation and horizontal gene transfer. Of particular interest was the status of clonal bacteria such as *M. tuberculosis*. I have showed that despite its limited diversity, variation exists within the genomes of this species, and that its potential to adapt must not be underestimated.

Cole ST, Brosch R, Parkhill J, et al (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. 393:537–. doi: 10.1038/31159

Colwell RR (1996) Global climate and infectious disease: The cholera paradigm. Science 274:2025–2031.

Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–1266. doi: 10.1534/genetics.106.063305

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351. doi: 10.1038/nrg.2016.49

Lecompte O, Ripp R, Thierry J-C, et al (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res 30:5382–5390.

McLafferty MA, Harcus DR, Hewlett EL (1988) Nucleotide sequence and characterization of a repetitive DNA element from the genome of Bordetella pertussis with characteristics of an insertion sequence. J Gen Microbiol 134:2297–2306.

Nikolayevskyy V, Kranzer K, Niemann S, Drobniewski F (2016) Whole genome sequencing of Mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: A systematic review. Tuberculosis (Edinb) 98:77–85. doi: 10.1016/j.tube.2016.02.009