

**Imperial College
London**

Genomic and phylogenetic analysis of infectious diseases

Arturo Torres Ortiz

Department of Infectious Disease
Imperial College London
a.torres-ortiz18@imperial.ac.uk

Supervised by:

Dr. Louis Grandjean, Prof. Xavier Didelot & Prof. Onn Min Kon

This thesis is submitted for the degree of
Doctor of Philosophy

October 2022

Declaration of Originality

I hereby declare that the work presented in this thesis is entirely my own. All other work is appropriately referenced in the text.

The contents of chapter 3 have been published in the following article:

Torres Ortiz A, Coronel J, Vidal JR, Bonilla C, Moore DAJ, Gilman RH, Baloux F, Kon OM, Didelot X, and Grandjean L (2021). Genomic signatures of pre-resistance in *Mycobacterium tuberculosis*. *Nature communications* 12(1): 7312.11

Chapters chapter 2 and chapter 4 are currently pending publication. The contents of chapter 4 can be found in the following article:

Torres Ortiz A, Kendall M, Storey N, Hatcher J, Dunn H, Roy S, Williams R, Williams C, Goldstein RA, Didelot X, et al. (2022). Within-host diversity improves phylogenetic and transmission reconstruction of SARS-CoV-2 outbreaks. *bioRxiv*: 2022.06.07.495142.11

Arturo Torres Ortiz
October 2022

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

First and foremost, I would like to dedicate this thesis to all the participants of the studies that are part of this work. I sincerely hope I have shed some light on the maladies that afflict you.

Many people have, in one way or another, contributed for me to be able to do the research described in this thesis. I am thankful to my supervisor, Louis Grandjean. Many things you have taught me are not in this thesis or in any of our published work, but I will carry them with me. Xavier Didelot has always been available with a rare combination of intellect and generosity. I am also thankful to Onn Min Kon, Francois Balloux and Lucy van Dorp for providing me with an academic home and plenty of stimulating conversations.

My deepest thanks are dedicated to my mother. Any difficulties I encountered during my doctoral work pale in comparison with the hardships and challenges you overcame to provide us with our education. To my father and sister, I will always be grateful to you. To Ilsa, your friendship is one of the most valuable things I have gained from my PhD.

Finally, to Lisa, my girlfriend when I started my doctoral work and my wife when I finished it. I will happily spend the rest of my life making up for the time we had to be apart. Thank you for your unwavering support and your encouragement. You have endured this as much or more than I did. Thanks are far too small for what I owe you in return.

Abstract

Infectious diseases pose an unsustainable human and economic burden, and despite control efforts they remain a leading cause of mortality and morbidity, last evidenced by the COVID-19 pandemic and its devastating impact all over the world. Effective control of infectious diseases requires the continuous monitoring and surveillance of outbreaks of emerging, re-emerging or endemic pathogens. Disease prevention strategies are further challenged by the increasing incidence of antimicrobial resistance, which has emerged as a risk to global health and threatens the effectiveness of public health interventions. Advances in our understanding of pathogen biology and the development of clinical microbiology methodologies are necessary to tackle the many challenges posed by infectious diseases. Recent years have witnessed rapid advancements in pathogen genomics that have transformed clinical microbiology, patient care and disease control efforts. Despite this, genome sequencing has yet to be completely implemented in routine microbiology.

This thesis aims to extend and facilitate the use of genomics in clinical microbiology, as well as to improve our knowledge of pathogen biology and evolution. First, I present an approach to reduce mapping bias in short-read sequencing. The proposed methodology increases accuracy in hypervariable genomic regions and reduces inaccuracies due to reference bias. Next, I define the concept of pre-resistance to describe the genetic predisposition of drug susceptible pathogens to acquire drug resistance in the future. I then show evidence of pre-resistance in *Mycobacterium tuberculosis*, demonstrating a differential risk of drug resistance acquisition among lineages and in mono-resistance strains. Furthermore, I identify candidate loci and genomic polymorphisms associated with a higher risk of resistance acquisition. Identifying markers of future antibiotic resistance could enable targeted therapy to prevent resistance emergence in *M. tuberculosis* and other pathogens. Finally, I describe the effects of within-host diversity in phylogenetic and transmission inference. Using simulations

and data from SARS-CoV-2 hospital outbreaks, I show that using within-host diversity improves phylogenetic inference and our understanding of who infected whom.

The results of this thesis show the potential of pathogen genomics within clinical microbiology, and highlight the opportunities to improve patient care and public health measures to reduce the human and social burden of infectious diseases.

Contents

1	Introduction	1
	Introduction to clinical microbiology	1
	Whole-genome sequencing technologies	3
	Applications of whole-genome sequencing in clinical microbiology	8
	Diagnostics and species identification	8
	Phenotype testing	11
	Pathogen typing and outbreak detection	14
	Phylogenetic approaches for infectious disease research	17
	Phylogenetic patterns of disease transmission	17
	Pathogen phylodynamics	20
	Transmission and surveillance	25
	Within-host diversity	26
	Within-host evolution of infectious diseases	26
	Effects of within-host diversity in clinical microbiology	32
	Genetics and microbiology of <i>M. tuberculosis</i>	34
	Genomic epidemiology during the COVID-19 pandemic	36
	Thesis statement	38
2	Reducing short-read mapping bias in <i>M. tuberculosis</i> using de-novo assembly	41
	Introduction	41
	Sample collection	43

Mapping reference bias in <i>M. tuberculosis</i>	44
Hybrid assembly and mapping of short-reads decreases mapping errors	47
Discussion	50
3 Genomic signatures of pre-resistance in <i>M. tuberculosis</i>	55
Introduction	55
Methods	58
Study design and sample selection	58
Whole-genome sequence analysis	59
Phylogenetic analyses	61
Time-to-event analysis	63
Genome-wide association of pre-resistance	63
Results	64
Population structure, genomic analysis, and patient demographics . .	64
Phylogenetic analysis and drug resistance emergence	67
The emergence of compensatory mutations	69
Phylogeographic history of <i>Mycobacterium tuberculosis</i> in Peru	72
Between lineage differences in drug resistance acquisition	75
The risk of developing MDR TB from isoniazid mono-resistance . . .	78
Genomic signatures of drug resistance acquisition	78
Discussion	83
4 Within-host diversity improves phylogenetic and transmission reconstruction of SARS-CoV-2 outbreaks	89
Introduction	89
Methods	92
Model for within-host diversity	92
Outbreak simulations	93
Amplification and whole-genome sequencing	94
Whole-genome sequence analysis of SARS-CoV-2 sequences	95

Results	96
Sampling, demographics and metadata	96
Genomic analysis of SARS-CoV-2 sequences	97
Within-sample variation is stable between technical replicates	97
Within-sample variation is shared between epidemiologically linked samples	100
Within-host diversity model outperforms the consensus model in simu- lations	102
Within-host diversity improves the resolution in SARS-CoV-2 phyloge- netics	105
Discussion	109
5 Discussion	115
Genomic analysis of pathogen sequences	115
Genetic markers of pre-resistance	117
Implications of within-host pathogen diversity	120
Future perspectives	123
Concluding remarks	126
References	129
A Appendix	163
Chapter 2	163
Chapter 3	172
Chapter 4	184

List of Figures

1.1	Mutation rate of pathogenic microbes	21
1.2	Measurably evolving populations	23
1.3	Within-host diversity and transmission	27
2.1	Overview of the two short-read pipelines	46
2.2	Mapping error depends on lineage and genomic structure	48
2.3	Phylogenetic inference for each workflow	49
3.1	Analysis of pre-resistance	58
3.2	Phylogenetic analysis of 3134 <i>M. tuberculosis</i> isolates	68
3.3	Inferred posterior density distribution of the earliest occurrence of resistance to first line antituberculous drugs	71
3.4	Dynamics of non-synonymous mutations in <i>rpoC</i>	73
3.5	Phylogeographic analysis of <i>M. tuberculosis</i> introductions to Peru	74
3.6	Hazard ratio and Kaplan-Meier curve for sublineages of <i>M. tuberculosis</i>	76
3.7	Hazard ratio and Kaplan-Meier curve for rifampicin acquisition	79
3.8	Genome-Wide association study (GWAS) results	81
4.1	Genomic analysis of technical duplicates	99
4.2	Probability of sharing within-host variants in sample pairs	102
4.3	Model of within-host diversity	103
4.4	Similarity scores for inferred trees	105
4.5	Inferred phylogenetic trees from outbreak simulations	106

4.6	Changes in temporal signal when using within-sample diversity . .	107
4.7	Within-sample variation improves resolution of infectious disease outbreaks	108
A3.1	Temporal distribution of included in the study	173
A3.2	Distribution of sampling dates by lineage	173
A3.3	Maximum likelihood phylogeny of the Samara (Russia) data set . .	174
A3.4	Temporal signal analysis	175
A3.5	Tree dating MCMC chain convergence	176
A3.6	Dynamics of non-synonymous mutations in the <i>rpoC</i> gene for Ser450Leu and non-Ser450Leu <i>rpoB</i> mutations	177
A3.7	Tree dating MCMC chain convergence for the global data set . . .	178
A3.8	Scaled Schoenfeld residuals	179
A3.9	Hazard ratio and Kaplan-Meier curve of 100 phylogenetic bootstrap replicates	180
A3.10	Hazard ratio and Kaplan-Meier curve for drug resistance acquisition in the population level dataset	181
A3.11	<i>Mycobacterium tuberculosis</i> lineage differences in non-bacterial factors	182
A3.12	Population correction for Genome-Wide association study (GWAS)	182
A3.13	Short-read alignment around GWAS polymorphisms	183
A4.1	Collection date distribution and time from symptom and days from symptom onset	184
A4.2	Sample mean coverage distribution	187
A4.3	Number of low frequency variants and C_t value	187
A4.4	Proportion of shared mixed variants between duplicated samples using different filters of allele frequency	188
A4.5	Probability that mixed variants are shared	189
A4.6	Phylogenetic trees for SARS-CoV-2	190
A4.7	Genetic distance between longitudinal samples	190

List of Figures

A4.8	Time calibrated phylogenetic trees for SARS-CoV-2	191
A4.9	Phylogenetic and transmission for SARS-CoV-2 outbreaks	191

List of Tables

1.1	Genome sequencing technologies	6
3.1	Population structure	66
3.2	First emergence of drug resistance conferring mutations	70
3.3	Gene-based association analysis	82
4.1	SARS-CoV-2 sample collection and demographics	98
4.2	Phylogenetic distance between pairs of samples	101
A2.1	Assembly metadata	163
A3.1	Drug resistance profiles by project	172
A4.1	Metrics used for phylogenetic tree comparison	184
A4.2	Transition/transversion rates and base frequencies of the known simulated tree	185
A4.3	Inferred transition/transversion rates and base frequencies when using the consensus sequence	185
A4.4	Inferred transition/transversion rates and base frequencies when accounting for within-host diversity	186

Acronyms

AMR Antimicrobial resistance. 2

DNA Deoxyribonucleic acid. 3

DR Drug-resistance. 36

DST Drug Susceptibility Testing. 2, 41, 59, 69, 84

EMB Ethambuto. 35

ESS Effective sample size. 68

GWAS Genome-wide association studies. 42, 80, 81, 86, 117, 118, 123, 124

HR Hazard Ratio. 75–80, 82

indel Insertions and deletions. 36, 60

INH Isoniazid. 35, 70

IQR Interquartile range. 67, 78, 96, 97, 100

L2 Lineage 2. 61, 65, 67, 68, 75, 78

L4 Lineage 4. 61, 65, 67, 68, 75, 78, 85

MALDI-TOF Matrix-assisted laser desorption/ionization time of flight mass spectrometry. 2, 9, 10

MCMC Markov Chain Monte Carlo. 62, 67, 68, 73

MDR Multidrug-resistance. 35, 36, 58

MIC Minimum inhibitory concentration. 3

ML Maximum likelihood. 67, 73

MODS Microscopic Observation Drug Susceptibility. 58, 59, 69

MRCA Most recent common ancestor. 23

MTBC *Mycobacterium tuberculosis* complex. 34, 35

NGS Next-Generation Sequencing. 4, 5, 7

ONT Oxford Nanopore Technologies. 5

PacBio Pacific Biosciences. 5, 44

PCR Polymerase chain reaction. 3, 9, 33, 94, 96, 99, 109

PE Proline-Glutamate. 61

PPE Proline-Proline-Glutamate. 61

PZA Pyrazinamide. 35, 70

R₀ Basic reproductive number. 24, 93

RIF Rifampicin. 35, 70

RNA Ribonucleic acid. 3

SMRT Single-molecule real-time sequencing. 5

SNP Single nucleotide polymorphism. 35, 36, 51, 60–66, 68, 80

STR streptomycin. 70

Acronyms

tMRCA Time to most recent common ancestor. 23, 68

VCF Variant Calling Format. 45, 46, 60

VoC Variants of concern. 37

WGS Whole-Genome Sequencing. 4

XDR Extensively drug-resistance. 35

Chapter 1

Introduction

More than 300 years ago, Antone Van Leeuwenhoek described the ubiquitous presence of microscopic organisms living around and within us. These organisms invisible to the naked eye are known as microorganisms or microbes. Since then, the essential role of microorganisms in many natural and industrial processes has been clearly demonstrated. Microbes constitute an indispensable part of many ecosystems and they form interdependent relationships with other species, including our own. However, some microbes are also causative agents of disease that have the potential to transmit among individuals or between individuals and their environment, thus their broad categorization as infectious diseases. In this section, I will focus on the biology and evolution of pathogenic microorganisms, their role in pathogenesis and the use of microbiology methodologies to identify and manage infectious diseases.

Introduction to clinical microbiology

Clinical microbiology focuses on the isolation and characterization of infectious microorganisms in order to provide the necessary information to aid prevention, diagnosis and treatment of infectious diseases (Bachman and LeBar 2019). The management of pathogens has important consequences not only for patient recovery, but also at the community level as it enables detection and control of disease outbreaks.

Infections can be caused by four kinds of microorganisms: bacteria, fungi, viruses, and parasites. Given the significant diversity between and within these groups, clinical microbiology involves a wide range of complex and species-specific biological methodologies to ensure proper collection, isolation, detection, classification and diagnosis of infections. The scope of clinical microbiology ranges from individual patient management to population and epidemiological dynamics of infectious diseases (Didelot et al. 2012). The applications of clinical microbiology can be grouped in three categories: diagnostic microbiology and species identification, phenotype testing, and outbreak monitoring.

Diagnostic microbiology focuses on the collection, identification and classification of pathogens from clinical isolates. As mentioned above, most of the methodologies involved are group- and species-specific, and therefore out of the scope of this thesis. Most detection approaches rely on direct identification of the pathogen or diagnosis of a pathogen-specific host response (Cintrón et al. 2019). For instance, many bacteria can typically be detected after growth in culture, where a sample is introduced into media with nutrients such as solid agar or liquid broth. The type of media, nutrients and growth type often depends on the species under consideration. Species are then identified by direct microscopy, gram staining, colony morphology, colorimetry, or MALDI-TOF mass spectrometry. Unculturable bacteria, viruses, fungi and eukaryotic parasites can be detected using serological, antigen and molecular assays (Bachman and LeBar 2019).

Once a pathogen has been detected, phenotype testing can provide clinically actionable insights for individual-patient and public health management, and therefore forms a critical part of clinical microbiology. The main microbial phenotypes of clinical interest are antimicrobial resistance (AMR) and pathogenicity (Balloux et al. 2018). Antimicrobial susceptibility testing is a critical part of patient management. Drug susceptibility testing (DST) is largely based on phenotypic testing, where pathogen

growth is detected in the presence of varying concentrations of antimicrobial. The concentration at which there is no growth is the minimum inhibitory concentration (MIC). Detection of virulence factors is rarely performed, but there are a few instances when it is routinely implemented using bioassays (analysis of effect on living organisms), serotyping (detection of pathogen-specific antigens) or molecular methods such as PCR (Didelot et al. 2012).

Public health microbiology is concerned with the epidemiology and ecology of infectious diseases. Typically, it involves detection and monitoring of disease outbreaks, as well as pathogen surveillance. Given that broad species determination is often not enough for epidemiological analysis, strain identification is carried out by typing methods. For example, *M. tuberculosis* typing is based on spoligotyping, where a highly polymorphic direct repeat locus is amplified by PCR (Tagini and Greub 2017).

Whole-genome sequencing technologies

All living cells store their hereditary information in the form of DNA (Deoxyribonucleic acid) or RNA (Ribonucleic acid), both formed of chains of four nucleotide bases that carry the information required to synthesise all of the organism's proteins (Alberts et al. 2022). The four bases in DNA are guanine (G), adenine (A), cytosine (C) and thymine (T); while RNA consists of uracil (U) instead of thymine. The complete set of genetic information in an organism or cell's DNA or RNA is called the genome (Giani et al. 2020). For over 50 years, multiple technologies have been developed to accurately determine the sequence of nucleotides that form a segment or the entirety of the genome.

Initial sequencing advances relied in the chain termination sequencing method published by Sanger et al. (1977), and therefore they are collectively known as Sanger sequencing, with modifications to include fluorescent labelling of nucleotide bases

and capillary technologies to allow more automatic workflows. Early sequencing projects depended on restrictions maps in known regions, which made sequencing of longer chromosomes challenging. To solve this, Staden (1979) proposed a methodology, later known as "shotgun sequencing", where the DNA is randomly fragmented into smaller segments for sequencing, and then produce a large amount of redundant sequenced fragments, also called reads, to assemble them into longer sequences by searching for overlaps (Kwong et al. 2015). This concept paved the way for ambitious projects aiming at sequencing the entire genome —whole-genome sequencing (WGS)—, catalyzed the development of newer sequencing technologies, and established the capital importance of computational algorithms to assemble the sequencing reads and analyse genomic data.

The currently most widely used sequencing platforms are summarised in Table 1.1. The low-throughput and complex laboratory protocols characteristic of Sanger sequencing preclude its use for whole-genome sequencing projects, and it has been largely replaced by other methodologies. However, due to the long and highly accurate reads produced by Sanger sequencing, platforms based on this technology are still used to sequence specific fragments of the genome or to prepare good quality reference whole-genomes.

In the early 2000s, new technologies were developed to overcome the limitations of earlier methods by performing massively parallel short-read sequencing. These methodologies are collectively known as Next-Generation Sequencing (NGS) or Second-generation sequencing (Giani et al. 2020). As these technologies became more efficient and affordable, an extensive amount of large whole-genomes have been sequenced and assembled. Of all the commercialised NGS methods, Illumina sequencing is by far the most widely used. Illumina sequencing technologies use a sequencing-by-synthesis method that detects fluorescent signals emitted when a nucleotide is incorporated in the complementary DNA strand. Illumina has quickly

become the leading technology for genome sequencing due to its high accuracy, low error rate, high throughput and the decreasing low cost per sequenced base which favored the implementation of whole-genome sequencing methods in many areas of research and clinical laboratories. The use of Illumina platforms is also supported since several platforms have been released to the market with moderate (MiSeq), moderate-high (NextSeq) and high (HiSeq) throughput (Kwong et al. 2015), allowing for multiple applications at different cost. However, the short-reads from Illumina (100-300 nucleotide base pairs, bp) limit the interpretation of the results and the applicability of NGS.

The downstream analysis of sequence reads typically involves genome assembly or read mapping. For genome assembly, long sequences called contigs are generated by detecting overlapping fragments between the reads. Read mapping relies on the alignment of sequencing reads to a reference genome by estimating matches between the reads and the target sequence. Both genome assembly and read mapping are greatly affected by the length of the sequencing reads. Shorter reads reduce the ability to detect overlaps between sequencing reads. Similarly, short reads are more challenging to map to a reference genome as the probability of uniquely matching a read to the target sequence increases with longer reads. This issue is further exacerbated by the presence of duplicated and low complexity DNA regions in the genome, which decrease the probability of uniquely mapping reads with good quality.

Third-Generation sequencing technologies address these issues by producing reads of thousands of base pairs, significantly longer than those typically found in NGS methods. Third-Generation sequencing methods generally encompass those that do not require DNA amplification before sequencing, and it is mostly represented by two leading and fundamentally different technologies: single-molecule real-time sequencing (SMRT-seq) by Pacific Biosciences (PacBio) and nanopore sequencing by Oxford Nanopore Technologies (ONT). SMRT-seq leverages the high sensitivity of zero-mode

Table 1.1: Genome sequencing technologies.

Platform	Advantages	Disadvantages
First-generation sequencing		
Sanger sequencing	Mid-size reads (~700bp) Highly accurate	Low throughput Labor intensive
Next-generation sequencing		
Illumina	High sequence throughput Low error rate Low cost per base	Short reads (100-300bp) Long run time Requires amplification
Third-generation sequencing		
PacBio	Long reads (10–20 kbp) High quality Short execution time Detection of epigenetic markers Randomly distributed errors	Low throughput High cost per base High instrumental cost Cumbersome instrument
Nanopore	Long reads (10–20 kbp) Small and highly portable Low cost benchtop instruments Detection of epigenetic markers	High error rate Systematic sequencing errors High cost per bp

waveguides (ZMWs), which allows emitted fluorescence to be observed in small quantities, and permits detection of the fluorescence emitted in single-molecules and in real-time during DNA replication. PacBio is able to generate an average read length of 10,000bp (or 10Kbp), with many reads surpassing 50kbp, allowing for complete and circularized prokaryotic genomes. Although the average base quality of PacBio reads is low compared to NGS technologies, the error profiles are purely random, facilitating their correction by increasing the genome coverage (Giani et al. 2020). Newer developments, such as HiFi ("High Fidelity"), also aim at increasing the base quality. Finally, PacBio can detect epigenetic changes, such as DNA methylation, which are increasingly being recognised for their importance in genome regulation. However, PacBio has a high cost per base, while the increased cost of the equipment and the high size of the instrument prevent its implementation in low- and middle-income countries. Nanopore sequencing technology uses a genetically modified bacterial nanopore placed in a sensor chip that measures the current running through the pore. As single DNA strands travel through the pore, the current is disturbed. Each nucleotide base changes the current in a different way, which is measured by the sensor, and the pattern of change is translated into a DNA sequence. Similarly to PacBio, Nanopore sequencing can produce reads with an average length of 10-20kbp and can detect epigenetic markers. Furthermore, many Nanopore sequencing devices are highly portable and can be as small as a USB device, facilitating sequencing in real time in low- and middle-income countries and remote regions (Quick et al. 2016). However, Nanopore presents a high error rate compared to NGS, and unlike PacBio, the sequencing errors are systematic, which makes it challenging to overcome even at high sequencing coverage. New developments and improvements in both the algorithm and the sequencing chemistry, such as the R12 flowcells or Duplex Sequencing, aim at improving the base quality, and the real-time results opens opportunities not available in other sequencing technologies (Payne et al. 2021).

Applications of whole-genome sequencing in clinical microbiology

The development of whole-genome sequencing technologies has rapidly placed pathogen genomics as a serious candidate to become standard practice in clinical microbiology. In principle, whole-genome sequences provide vast amounts of information that can readily complement or even replace many of the routine clinical microbiology tests. First, genome sequences are especially well suited to characterize microbial species and strains at a higher speed and resolution than classic microbiology assays. Moreover, genetic data can be used to predict pathogen phenotypes of clinical importance. Over the last decade, researchers have identified genetic determinants of many pathogen phenotypes that influence clinical decisions and patient outcome, particularly antimicrobial resistance and, to a lesser extent, pathogenicity. Last, genome sequences allows detection and monitoring of outbreaks with high resolution to aid epidemiological and public health practice. Crucially, the promising outlook of whole-genome sequencing in clinical microbiology relies on providing all these information and applications within a single and relatively straightforward assay instead of using multiple technologies (Didelot et al. 2012).

Diagnostics and species identification

Rapid pathogen species characterization is a critical part of clinical microbiology as it provides relevant information about prognosis and disease management (Köser et al. 2012). Identification of the pathogen species present in a clinical sample is often the first step in the clinical microbiology pipeline, guiding and affecting most of the subsequent procedures. Whole-genome sequencing is well suited to rapidly identify pathogen species, providing a much higher resolution than existing diagnostic methods.

Diagnostic microbiology has greatly benefited from many technological advancements, resulting in a plethora of methods available for clinical diagnosis. Most applications in clinical microbiology can be largely grouped in four categories: pathogen culture, direct microscopic examination, serological assays, and molecular diagnostics (Cintrón et al. 2019). Microbe identification methods based on antigenic detection and polymerase chain reaction (PCR) methods have greatly reduced complexity, cost and turnaround time, increase service and expanding capabilities beyond specialized reference laboratories (Murray et al. 2021). Notably, the development of MALDI-TOF MS (matrix-assisted laser desorption/ionization time of flight mass spectrometry) has revolutionized diagnostic microbiology. Due to its rapid, consistent, accurate and highly cost effective identification of microbial pathogens, MALDI-TOF has become the predominant method and the gold standard for microbial identification in many laboratories (Cintrón et al. 2019).

At the broadest level, genome sequences from clinical isolates can be rapidly compared to existing reference databases in order to established their taxonomic classification. Accurate species identification can be achieved either by comparing whole-genome sequences to a catalog of reference whole-genomes or by using targeted approaches to sequence specific genes such as 16S rRNA or multi-locus sequence typing (MLST), therefore increasing diagnostic speed and reducing sequencing costs (Bertelli and Greub 2013). Moreover, genome sequences add an evolutionary perspective to pathogen taxonomy, allowing for a natural classification system that will change current taxonomic groupings and simplify diagnosis (Didelot et al. 2012). Sequencing data offers higher resolution than other diagnostic methods, it is less susceptible to convergent evolution (which affects evolutionary analyses), and it facilitates comparative analysis. As a result, genome sequence data has shown that many long-established pathogen classifications do not represent monophyletic groups and it has allowed for the separation of closely related species (Didelot et al. 2012). This consistent and robust classification system will clarify pathogen diagnosis, which in

turn will facilitate the identification of emerging or novel pathogens for which no direct reference database exists (Bertelli and Greub 2013).

However, the use of whole-genome sequences to exclusively identify pathogen species alone may not be justified for routine clinical microbiology given the low cost of current diagnostic methodologies, especially MALDI-TOF. Nevertheless, the promise of genomics relies on being an all-in-one method that can additionally deliver information regarding strain typing or phenotype testing, as later discussed. Moreover, there are still multiple potential applications of whole-genome sequencing within diagnostic microbiology. MALDI-TOF has shown great accuracy at the genus level, but variable results when identifying pathogens at the species level (Hou et al. 2019). The high accuracy of whole-genome sequencing can be leveraged when species level resolution is required and other methods show conflicting results. Genome sequencing can also be used when other methods consistently misdiagnose, such in the case of novel pathogens or hard to identify variants and species (Köser et al. 2012). Another advantage of sequencing is that it allows diagnosis of complex mixtures of pathogens directly from clinical samples, which is still under development for MALDI-TOF (Mahé et al. 2014; Yang et al. 2018). Metagenomic approaches that analyze the population composition of clinical samples have shown great sensitivity (Charalampous et al. 2019), promising an unbiased and untargeted diagnostic method. Sequencing directly from clinical samples can also be advantageous for the detection of non-culturable or slow-growing pathogens. Moreover, some sequencing platforms such as the MiniON system (Nanopore, Oxford, United Kingdom) also offer low-cost, real time sequencing with unparalleled portability directly from the clinical sample (Hasman et al. 2014), which facilitates its implementation in settings with minimal laboratory capacity (Gardy and Loman 2018). Whole-genome sequences also stand out in the identification of species with large accessory genomes such as plasmids, which encode large degrees of genetic diversity, including pathogenicity and virulence plasmids of clinical relevance that will be missed with other diagnostic methods. The cost of

whole-genome sequencing can also be lowered since the identification of most species do not require a complete, high-quality assembly. Genome finishing and gap closure is expensive and time-consuming, but is often unnecessary for species identification and other clinical microbiology approaches since incomplete and even low coverage samples are sufficiently informative (Greub et al. 2009).

Phenotype testing

Pathogen characteristics of clinical relevance may affect patient's clinical management and outcome, especially phenotypes related to anti-microbial resistance and pathogenicity. In principle, whole-genome sequences provide with information that can be used to understand the genetic basis of phenotypes of interest and to develop genetic markers for diagnosis.

Drug susceptibility detection is key in the management of infectious diseases, both at the patient and public health level. The success at which genetic markers of antimicrobial resistance have been identified varies by species and genomic mechanism underlying the phenotype. Whole-genome sequencing is very well suited to detect the presence of genetic elements that confer drug resistance, such as the insertion of the *SCCmec* cassette in *Staphylococcus aureus* carrying the gene *mecA*, which confers methicillin resistance (Matsushashi et al. 1986; Bode et al. 2012). Transferable plasmid-mediated carbapenemase-encoding genes among gram-negative pathogens, especially Enterobacteriaceae, also represents a good predictor of carbapenem-resistance, and has been successfully implemented as a genetic marker in clinical settings (Codjoe and Donkor 2017). As our knowledge of antimicrobial resistance associated genes grows, susceptibility testing using whole-genome sequencing will be more accurate. For instance, there are currently over 3,000 resistance genes recorded in the ResFinder database (Florensa et al. 2022). Mutations in existing genes can also confer resistance to antimicrobials, and these mutations have successfully been implemented

as markers of resistance. Point mutations that confer drug resistance to first-line antibiotics such as isoniazid, rifampicin, ethambutol and streptomycin have been identified in *M. tuberculosis*, and they have been applied to infer drug-resistance phenotypes with a sensitivity and specificity similar or superior to that of laboratory drug susceptibility testing (The CRyPTIC Consortium and 100000 Genomes Project 2018).

Similarly to species identification in metagenomics, genomic data also facilitates the discovery of genetic markers of antibiotic resistance directly from clinical or environmental samples. The realization that soil samples contained an important reservoir of drug resistance genetic determinants prompted D’Costa et al. (2006) to coin the term “resistome”, later defined as the overall collection of all the antibiotic resistance genes and their precursors in bacteria (Wright 2007). The impact of the resistome in clinical microbiology is still unclear, but recent studies have suggested its important role in the evolution, maintenance and transmission of drug resistance genes (Kim and Cha 2021). Accurately characterising the resistome using genome sequencing could provide an unbiased diagnosis of all resistance genes within a clinical sample to guide clinical treatment

However, whole-genome sequencing is currently unable to reliably predict drug resistance acquired by other genetic mechanisms. Changes in gene expression can confer antimicrobial resistance, and such changes can only be predicted once the association between DNA mutations and gene expression has been clearly established using RNA sequencing methods. For instance, *ampC* overexpression confers beta-lactam resistance in *Escherichia coli* (Caroff et al. 2000), or mutations in complex regulatory systems cause vancomycin resistance in *S. aureus* (Mwangi et al. 2007; Cameron et al. 2016). Moreover, drug-resistance resulting from intrinsic and non-specific mechanisms such as membrane permeability or modifications in the expression or structure of efflux pumps is currently challenging to infer (Ellington et al. 2017).

Genome sequences can also be used to characterize genetic markers of virulence, which encompasses many traits of interest such as disease severity, clinical presentations, immunogenicity or transmission potential. A clear example of differential virulence is found in *Neisseria meningitidis*, where the presence of the bacteriophage MDA Φ integrated into the *N. meningitidis* genome was associated with invasive infections and meningitis disease (Bille et al. 2005). Many genetic markers of virulence include toxins such as the Panton-Valentine leukocidin (PVL) phage-encoded toxin in *S. aureus* associated with necrotising pneumonia (Tong et al. 2015), or the Shiga toxins (verotoxins) in *E. coli* (Robins-Browne et al. 2016). Other genes associated with virulence are those related to biofilm or surface adherence in *S. aureus* (Tagini and Greub 2017). Genomic studies in *M. tuberculosis* have identified a wide range of virulence mechanisms that the bacilli employ to survive and persist in the hosts, including lipid and fatty acid metabolism, cell wall proteins and lipoproteins, secretion systems such as the Type VII systems, kinases, proteases or host immune system interaction, to name a few (Forrellad et al. 2013).

Decreased immunogenicity contributes indirectly to virulence and has implications both for clinical patient management and public health control programs. Genomic sequences can identify genetic mechanisms of immune system evasion. For instance, some SARS-CoV-2 genetic mutations and variants present a decreased immune response when compared to strains circulating earlier during the coronavirus disease 2019 (COVID-19) pandemic (Planas et al. 2021). Similarly, genetic markers of differential transmission is of great relevance to public health and outbreak management. Variants of SARS-CoV-2 have been linked to an increase in transmission potential, which affects epidemiological estimates and disease control measures (Davies et al. 2021a).

Pathogenicity based on broader lineage genetic markers also holds promise. For

instance, given that *M. tuberculosis* has a slow substitution rate and it is mainly clonal, describing virulence genes or mutations is often challenging. However, some lineages of *M. tuberculosis* are inferred to be more virulent than others (Ford et al. 2013), even though the specific mutations and molecular mechanisms for that different virulence remains unknown.

However, the understanding of the genetic determinants of virulence is still insufficient to translate into clinical and public health practice. Prediction of virulence phenotypes based on genomic data is challenging since many clinical presentations also depend on multiple host factors, patient comorbidities and environmental variables. Moreover, virulence is often determined by differential expression of virulent genes and therefore genetic sequences alone are insufficient to characterize it. Finally, virulence itself is a complex phenotype and its definition may vary among pathogens.

Pathogen typing and outbreak detection

Whole-genome sequences can be leveraged to identify and relate organisms with an unparalleled and an unprecedented resolution (Didelot et al. 2012). One immediate application is high-resolution pathogen typing for infection control and outbreak investigation. Pathogens that are part of the same outbreak or transmission chain are expected to be more similar to each other than to those that are not epidemiologically related. Many classical typing methods lack enough resolution to distinguish closely related isolates such as those that are part of an outbreak or a transmission chain. In contrast, genome sequences are characterized by higher discriminatory capacity which can be exploited for real-time surveillance of infectious diseases. Moreover, whole-genome sequencing can be applied regardless of the organism, thus having the potential to replace complex and numerous species-specific typing methods.

Retrospective studies have shown the potential of whole-genome sequencing in

reconstructing disease outbreaks. An early application of genomics for molecular epidemiology by Harris et al. (2013) rapidly described a detailed hospital and community outbreak of methicillin-resistant *S. aureus*, showing the improvement over conventional infection-control methods. Moreover, a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* was investigated using whole-genome sequencing, identifying one single origin for the outbreak with multiple initial transmission events, including a contaminated ventilator (Snitkin et al. 2012). Similarly, (Houlihan et al. 2018) described infectious disease outbreaks at high-resolution using whole-genome sequencing in influenza. Whole-genome sequencing also revealed frequent transmission of *Mycobacterium abscessus* among patients with cystic fibrosis, calling for additional measures to manage such patients (Bryant et al. 2013), and aided inference of risk factors for *M. tuberculosis* transmission in an outbreak among drug users in British Columbia, where it proved several independent transmission events as opposed to the single outbreak suggested by classic typing methods (Gardy et al. 2011).

The high-resolution provided by whole-genome sequencing technologies can enable real-time surveillance of infectious disease outbreaks, as seen in field genomic tracking of ebola outbreaks in Africa (Quick et al. 2016) and arbovirus such as dengue (Adelino et al. 2021) and zika virus (Faria et al. 2017) in the Americas. A German *E. coli* outbreak in 2011 became an early example of how genomic data can be leveraged to rapidly monitor disease outbreaks. Within days of the identified onset, a novel *E. coli* O104:H4 was identified as the cause of the outbreak (Mellmann et al. 2011).

The COVID-19 pandemic represents a stepping stone in the use of real-time genomic surveillance to support public health decision-making, both at a local and international scale. Rapid SARS-CoV-2 sequencing led to real-time review of infection control procedures in a renal unit at a UK hospital where results were reported and actions were enacted within the timeframe of the ongoing outbreak (Meredith et al. 2020). Similar patient management changes were established in school settings (Ismail et al.

2021), a care home in the UK (Ladhani et al. 2020) and an acute care hospital in Boston (Klompas et al. 2021), where rapid genome sequencing enabled to identify the origin of the outbreak, define risk factors of infection and implement targeted measures to disease control.

Apart from allowing pathogen typing based on sequence similarity, whole-genome sequences can be placed within a phylogenetic context, providing with a consistent and rigorous pathogen classification system. For instance, highly clonal species such as *M. tuberculosis* are well suited for a robust phylogenetic classification which can be reduced to minimal sets of mutations that can rapidly differentiate lineages and sublineages with great detail (Coll et al. 2012). The use of a phylogenetic framework has also been essential in the classification of SARS-CoV-2 variants, providing a dynamic system necessary to easily couple the emerging diversity arising during an ongoing pandemic (Rambaut et al. 2020). Furthermore, the inherent evolutionary element in this classification systems allows further linking of clustered groups with phenotypes of clinical interest such as the variants of concern for SARS-CoV-2 (Attwood et al. 2022) or antigen diversification for the periodical influenza vaccine update (Steinbrück et al. 2014).

The added value of whole-genome sequencing to epidemiology and public health has been repeatedly proven (Didelot et al. 2012). However, the use of genomics in pathogen typing and outbreak tracking depends on the genomic variability between organisms and the use of a molecular clock, which describes the constant changes that accumulate over time in genetic sequences (Campbell et al. 2018). That is why most examples of real-time disease surveillance have been in viruses, which are characterized by a substitution rate orders of magnitude higher than that of bacteria. Despite this, genome-sequencing real-time surveillance has shown applicability potential in many bacterial pathogens such as *S. aureus* and *Clostridium difficile* (Eyre et al. 2012). More importantly, even when transmission chains cannot be precisely reconstructed,

genomic sequences can still be leveraged to infer outbreak structure, transmission patterns and to describe risk factors of infection.

Phylogenetic approaches for infectious disease research

As mentioned above, genomic sequences and genetic distances can be used to cluster pathogens for outbreak analysis and identification of linked transmission cases. More importantly, genetic sequences reflect the evolutionary history of pathogens, and therefore they can be analyzed using evolutionary and phylogenetic models. Given the fast rate of evolution of many pathogens, as well as their large population sizes and short generation times, evolutionary timescales can often be interpreted within ecological and outbreak timeframes (Pybus and Rambaut 2009). Therefore, the relatedness of genetic sequences reflect epidemiological relationships, and phylogenetic models can be leveraged to infer parameters of clinical and public health interest (Campbell et al. 2018).

Phylogenetic patterns of disease transmission

At the broadest level, phylogenetic analysis provide information regarding the structure of the genetic diversity among pathogen isolates. Ultimately, phylogenetic structure represents variation in evolutionary history among isolates, which provides useful insights from a clinical and public health perspective. For instance, the distribution of different isolates along a phylogenetic tree can help determine whether they belong to the same outbreak or to completely different transmission events (Lemieux et al. 2021). If pathogen isolates with epidemiological links cluster far apart in the phylogenetic tree, it can cast doubts on their potential epidemiological relationship.

Phylogeographic approaches combine the phylogenetic structure with geographical data, which can help determine the geographical origin and transmission patterns of disease outbreaks. During the beginning of the SARS-CoV-2 pandemic, the phylogenetic structure of viral isolates help determined the existence of multiple introductions in many regions of the United States (Bedford et al. 2020; Deng et al. 2020; Gonzalez-Reiche et al. 2020), as well as to distinguish importation events from cryptic continuous transmission in Washington State (Bedford et al. 2020). Understanding whether new detected cases are caused by importations or ongoing outbreaks is of high relevance to establish effective control measures. Importantly, shifting patterns in phylogenetic structure over time can also reveal the effects of public health interventions and thus aid in the assessment of the efficacy of control efforts.

The structure of the tree also resembles the evolution and epidemiology of pathogen phenotypes of clinical interest. Drug resistance in *M. tuberculosis* poses a serious threat to tuberculosis treatment and control. Acquired drug resistance due to treatment failure has traditionally been thought to be the primary mechanism of drug-resistance emergence (Ershova et al. 2014). However, the role of transmission of already drug resistant strains is increasingly being recognized (Yang et al. 2017). Distinguishing between acquired and transmitted drug resistance is challenging. A history of previous antituberculous treatment is often used as a proxy for acquired drug resistance (Odone et al. 2016), even though previous treatment has been identified as a poor predictor for acquired drug resistance (Andrews et al. 2010). The distinction between both causes of drug resistance emergence in tuberculosis is important to effectively address the drug resistant tuberculosis epidemic. If acquired drug resistance is the leading mechanism of drug resistance acquisition in a specific setting, effective treatment should be the priority to avoid the emergence of resistance. But if transmitted drug resistance is identified as a relevant cause in the expansion of drug resistance, then transmission control measures need to be included in the disease management strategies (Nsofor et al. 2017). The phylogenetic structure of tuberculosis isolates can

aid distinguishing acquired and transmitted drug resistance. If drug resistant isolates are distant and scattered along the phylogenetic tree, those isolates are possibly the result of independent acquisition of resistance, possibly during treatment. On the contrary, clustering of drug resistant isolates in the phylogeny can be an indication of transmission of drug resistance bacteria.

The distribution and structure of genetic diversity among the phylogenetic tree can also reveal patterns of selection that are clinically relevant. Recurrent mutations along the phylogenetic tree without a common ancestral origin are referred to as homoplasies, and can be indicative of selection. Many phenotypes under selection are clinically important, such as virulence, transmissibility or drug resistance. Phylogenetic trees and the identification of homoplasies can be exploited to identify relevant mutations. For instance, homoplastic mutations in *M. tuberculosis* conferring drug resistance were identified using a phylogenetic approach designed to detect homoplasies (Farhat et al. 2013). The identification of homoplastic mutations in SARS-CoV-2 has also been leveraged to identify variants of concern and candidates for genetic polymorphisms that can potentially confer an adaptive advantage, spread faster among the population and have clinical and public health consequences (Dorp et al. 2020; Volz et al. 2021a). For instance, the spike protein deletion Δ H69/V70 has arisen independently in multiple lineages and variants of concern, and was found to increase infectivity by enhancing incorporation of cleaved spike into virions (Meng et al. 2021).

The fine structure of a phylogenetic tree can also be indicative of evolutionary processes of epidemiological relevance. Mutations are the basis upon which phylogenies are inferred, and given that evolutionary and epidemiological timeframes overlap in pathogen phylogenetics, mutations will accumulate in different patterns depending on the transmission network. This realization means that the tree shape can be indicative of the underlying transmission patterns, and that the topological properties

of a phylogeny can help distinguish infectious disease outbreaks caused by different epidemiological processes such as super-spreader events or homogeneous transmission (Colijn and Gardy 2014).

Pathogen phylodynamics

The intricate relationship between the epidemiology and the evolutionary history of pathogens has led to the development of the field of phylodynamics (Grenfell et al. 2004). Phylodynamics focuses on the study of how the combination of host immune selective pressures, epidemiology and evolutionary processes shapes and influences pathogen phylogenies (Volz et al. 2013). Using viral genetic sequences, Grenfell et al. (2004) showed that changes in the phylogenetic structure over time are determined by variable pathogen population size, transmission patterns, immune selection and within-host population dynamics. Given the high substitution rate of RNA viruses, it is not surprising that Grenfell et al. (2004) focused on viral genomes. Bacteria, for example, are characterized by a mutation rate several orders of magnitude lower than that of viruses (Fig. 1.1a). However, given their larger genomes and the overall inverse correlation between mutation rate and genome size, the genome-wide mutational landscape of bacteria is comparable to that of many viruses (Fig. 1.1b).

Phylodynamic approaches integrate phylogenetic models, such as the coalescent or birth-death, with mathematical and epidemiological modelling of infectious diseases to incorporate data from several sources and infer parameters of epidemiological interest. Additionally, a phylogenetic frame also allows to investigate past population dynamics. Gire et al. (2014) employed phylodynamic methods to infer the early origin and spread of ebola virus from a single zoonotic spillover, dating the origin of the outbreak in Guinea in February 2014 and its spread to Sierra Leone by April 2014, while Dudas et al. (2017) used 1,600 ebola genomes to reconstruct the temporal and

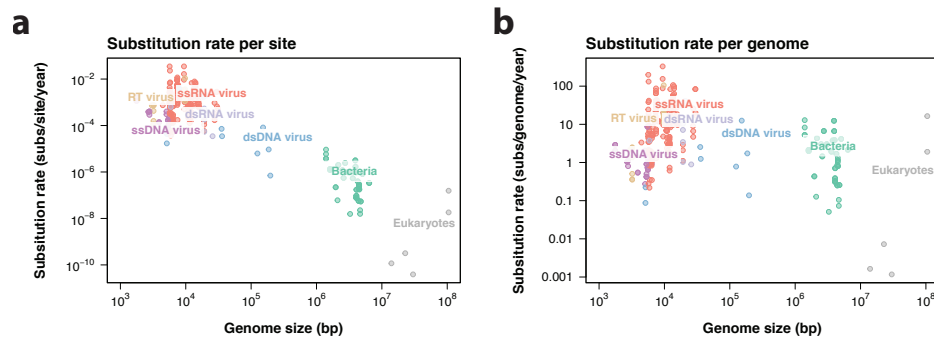


Figure 1.1: Mutation rate of pathogenic microbes.

Substitution rate of pathogenic microbes as a function of genome size. Colors represent major groups of pathogen classification: Eukaryota, bacteria, double-stranded DNA virus (dsDNA), single-stranded DNA virus (ssDNA), double-stranded RNA virus (dsRNA), single-stranded RNA virus (ssRNA), and reverse-transcribing viruses (RT). Labels are positioned in the mean substitution rate and genome size for each group, with a slight jitter added for representation. **a** Mutation rate calculated as substitution per site per year. **b** Mutation rate calculated as substitution per genome per year. Estimates of the substitution rate were obtained from Sanjuán (2012), Duchêne et al. (2016), and Duchêne and Holmes (2018).

geographical spread of ebola virus across West Africa and reveal important drivers for its spread. Similarly, Faria et al. (2017) and Metsky et al. (2017) inferred the transmission dynamics of Zika virus in the Americas, revealing Brazil as the outbreak's source to the rest of the continent.

A key element to phylodynamic studies is the time-calibrated phylogeny, a rooted phylogenetic tree with branch lengths measured in unit of time, nodes aligned to estimated dates of divergence and a per-year substitution rate. The most common source of information for phylogenetic pathogen dating methods is the sampling dates for the pathogen isolates. Therefore, in a phylogenetic tree calibrated using isolate collection dates, the leaves of the tree will also align at known dates of isolation (Didelot et al. 2018). Time-calibrating a phylogenetic tree allows to measure the evolution of an outbreak through time, infer transmission pairs and estimate the time of specific events of epidemiological importance. All dating methods rely on the assumption of a "molecular clock", which establishes that mutations accumulate steadily at a

particular rate over time (Kimura 1987), hereby connecting the temporal information within the sampling framework of an outbreak to the genetic diversity of the genomic sequences. Even though the assumption of a constant evolutionary rate has paved the way for phylodynamic approaches, evidence shows that it is a simplification as evolutionary rates vary within genetic sequences and between lineages of the same species (Welch and Bromham 2005), violating the assumption of a fixed and constant rate of evolution (Britten 1986; Ayala 1997; Jenkins et al. 2002). To overcome these limitations, several methods have been developed to time-calibrate the phylogenetic tree allowing for a variable rate of evolution ("relaxed clock") between lineages in a local clock approach (Yoder and Yang 2000), through child branches in an autocorrelated model (Thorne et al. 1998) or between all branches in uncorrelated models (Drummond et al. 2006).

The combination of epidemiological and evolutionary patterns require of measurably evolving populations and the accumulation of enough genetic diversity during the time of the sampling frame. Drummond et al. (2003) introduced the concept of measurably evolving populations to define those populations that exhibit detectable amounts of evolutionary change when sampled at different time points (temporal signal), which is a prerequisite to dating phylogenetic trees (Fig. 1.2). Measurably evolving populations accrue genetic diversity as time progresses, and thus a positive correlation is expected between the number of mutations from the ancestor of the population and the sampling dates in serially sampled isolates. This can be observed by fitting a linear regression between the distance to the root of the tree and the isolates sampling dates (Fig. 1.2). A more robust approach to understand the extent of observable evolution is the date-randomization test (Duchêne et al. 2015), which involves comparing parameter estimates, such as the substitution rate, obtained using the real data and multiple randomized datasets generated by permutation of sampling times. The randomized datasets represent the expected estimates in the absence of observable evolution (Rieux and Balloux 2016).

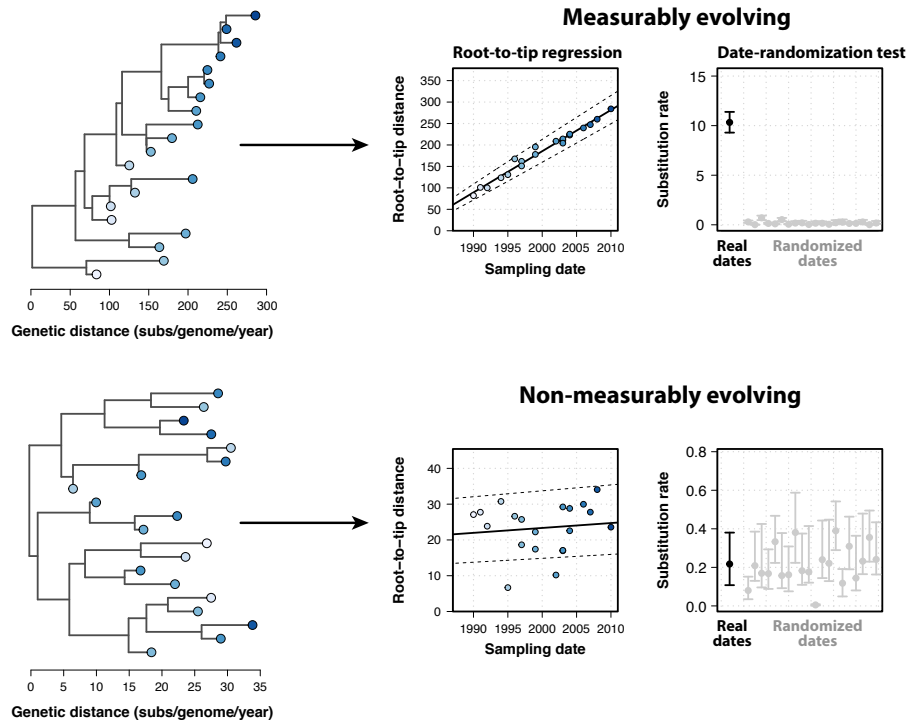


Figure 1.2: Measurably evolving populations.

Measurably evolving populations exhibit significant genetic diversity observable in serially sampled isolates (temporal signal). Measurably evolving populations are expected to accrue genetic mutations as time progresses, which can be observed by fitting a linear regression between the distance of the tips to the root (root-to-tip distance) and the dates of isolation. Additionally, temporal signal can be tested by comparing parameter estimates (such as the substitution rate) between the original dataset and multiple randomized sets generated by permutation of the sampling times to verify that the estimates with real data are not contained within the estimates of the randomized sets (date-randomization test). The top phylogenetic tree consists of a measurably evolving population, while the bottom phylogeny lacks temporal signal. A gradient of colors is used for the sampling dates, with lighter colors representing earlier sampling dates.

Time-calibrated phylogenetic trees allow estimation of the timing of specific events of epidemiological importance, which align to the age on internal nodes in a phylogenetic tree. For instance, the root of the phylogenetic tree for the isolates of an outbreak represents the most recent common ancestor (MRCA), and therefore the time of the root (time to most common recent ancestor, tMRCA) aligns with the date

of origin of the outbreak. Inferring the origin of an infectious disease outbreak has interest not only from a historical perspective, but also to provide actionable insights for infection control. Infectious diseases spread rapidly and often undetected during the first months of an outbreak, sometimes until community transmission is widespread and more radical interventions are needed (Gardy and Loman 2018). Understanding how long disease outbreaks have gone undetected provides an important source of information to measure the impact of early detection measures.

The timing of internal nodes can also provide insights into the evolution of mutations and phenotypes of clinical interest. For instance, Eldholm et al. (2015) used *M. tuberculosis* isolates sampled over 13 years to estimate that the outbreak was continuously circulating for over 4 decades, and used phylogenetic approaches to reconstruct the dynamic of drug resistance acquisition and to show that extensively drug resistance strains were in circulation years before the outbreak was detected.

Moreover, time-calibrated phylogenies can be directly coupled to mathematical models to infer epidemiological parameters such as the pathogen effective population size or the basic reproductive number (R_0), as well as their variation through time and in response to public health interventions (Volz et al. 2020). Pathogen effective population size can be used as a proxy for disease case counts, and thus its variation is reflective of the impact of infection control strategies, as shown in the declining population size after restricting movement across nearby regions during the ebola outbreak (Dellicour et al. 2018). Similar methods were used by Faria et al. (2014) to infer the early expansion of HIV-1, its geographical origin and the effect of policy and social changes on its spread. The basic reproductive number R_0 represents the average number of secondary cases per index case. Therefore, an R_0 higher than 1 indicates a growing epidemic, while a receding outbreak will show an R_0 lower than 1. Estimates of R_0 can be obtained from phylogenetic trees (Stadler et al. 2012) as previously undertaken in ebola (Stadler et al. 2014), and their variation through time

gives an indication of the transmission dynamics in response to control efforts (Volz and Didelot 2018).

Transmission and surveillance

Phylogenetic methods and genetic data have been recently used to resolve the order and timing of transmission events. A transmission tree is a graph representation of who infected whom in a transmission chain. Epidemiological data has been used to infer transmission trees, including contact tracing information or date of symptom onset, but often this type of data is not available or it is unreliable. Genome similarity and phylogenetic distance among pathogen isolates provides useful information regarding transmission events, under the assumption that pathogens will cluster together in the phylogeny in instances of direct transmission. Even though the structure of a phylogenetic tree may resemble and be affected by that of the transmission tree, both objects are not interchangeable (Pybus and Rambaut 2009). Phylogenetic trees represent the order and timing of coalescent events at the nodes of the tree, which occur prior to transmission, and with all sampled pathogen isolates at the leaves even when an isolate originates from another sample through a transmission event. Therefore, transmission events can take place at any point along the branches of the phylogenetic tree (Didelot et al. 2014; Hall et al. 2015). In transmission trees, hosts are represented by nodes, which are connected to other nodes by edges that represent transmission events. However different to one another, phylogenetic trees can still be used to infer transmission chains and the likelihood of transmission events (Morelli et al. 2012).

Sequence similarity among pathogen isolates alone can suggest transmission events (Jombart et al. 2011), but often it is necessary to couple genetic and epidemiological data to accurately determine transmission dynamics. Initial attempts in inferring transmission trees using genomic data employed genetic distances among isolates

to complement epidemiological data (Cottam et al. 2008; Morelli et al. 2012; Ypma et al. 2012), but did not leverage the advantages of using evolutionary models and time-calibrated phylogenetic trees. Ypma et al. (2013) and Didelot et al. (2014) developed a model to infer transmission trees from dated phylogenetic trees. Both approaches modelled the evolution of populations within each host, explicitly separating branching and transmission events. This model was later extended by Didelot et al. (2017) to accommodate partially sampled outbreaks and unsampled intermediate links.

Within-host diversity

During the course of an infection, several genetically different populations of a pathogen coexist within the host. Due to its higher discriminatory power when compared to other typing methods, the use of whole-genome sequencing has paved the way to allow the study of microbial diversity within a single host. The extent of diversity accumulated depends of several factors, but it can range from relatively similar within-species heterogeneity to multi-species microbiomes. Within-host diversity is of critical importance to clinical microbiology, as it has the potential to affect every single part of the clinical microbiology pipeline, from species identification to epidemiology and control measures.

Within-host evolution of infectious diseases

Within-host diversity can be generated by three main mechanisms: *de novo* mutations within the host, co-transmission of a diverse inoculum, or superinfection (Fig. 1.3). First, populations may differentiate within the host by accumulating mutations, even if the original infection consisted of a single population. A co-transmission occurs when genetically distinct populations are transmitted in a single transmission event

from a donor harboring multiple populations. Last, a superinfection occurs when an individual is infected by two or more separate transmission events.

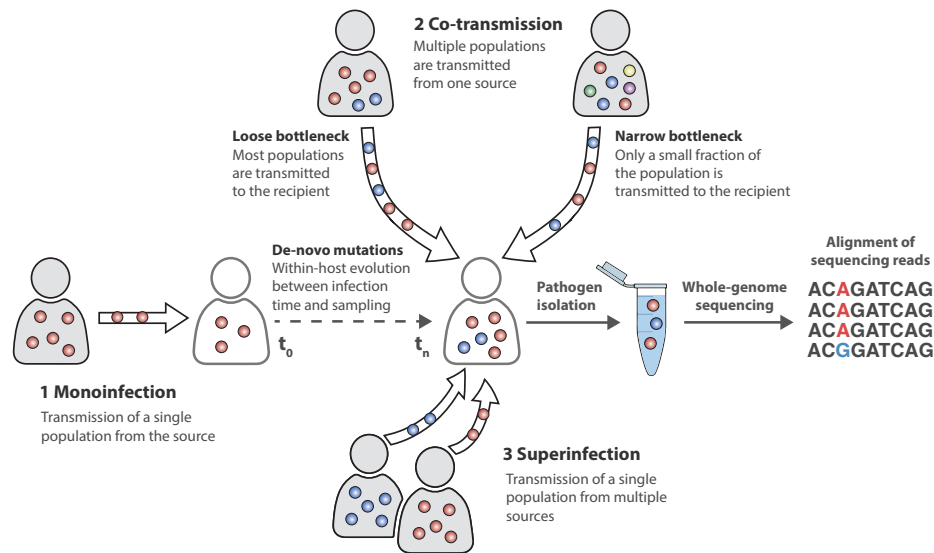


Figure 1.3: Within-host diversity and transmission of infectious diseases.

Within-host genetic diversity of infectious diseases is shaped by three different mechanisms. Mono-infection occurs when a single population is transmitted from the infection source to the recipient. Then, *de novo* mutations within the host can give rise to multiple subpopulations. In a co-transmission event, a recipient bearing multiple genetically distinct subpopulations transmits a diverse inoculum to the recipient. The extent of diversity transmitted will depend on the bottleneck size. Finally, a superinfection event is characterized by multiple infections from more than one source, each of them carrying a distinct population. Within-host diversity can be detected using whole-genome sequencing by aligning the reads and analyzing the diversity of nucleotide bases at each position.

As pathogens replicate during the course of an infection, mutations accumulate over time. *De novo* mutations provide the substrate for the evolution of within-host subpopulations, and the processes involved in the maintenance of this diversity depend on pathogen and host factors. The amount of diversity depends largely on the mutation rate. Therefore, rapidly evolving pathogens such as viruses are expected to harbor a higher degree of within-host diversity. Not surprisingly, much of the first evidence of the presence and clinical relevance of within-host diversity comes from rapidly

evolving RNA viruses such as HIV (Fraser et al. 2014). However, as mentioned in the previous section, analyzing the genome-wide mutational landscape of bacteria unravels similar levels of genetic diversity (Fig. 1.1). Importantly, estimates of mutation rates are largely based on molecular clock approaches that infer long-term mutation rates. Given the inverse relationship between mutation rate and the time frame of measurement, short-term pathogen mutation rates within the course of infection can far exceed those inferred from longer time frames (Ho et al. 2011). The apparent higher short-term evolutionary rate can be explained due to the presence of transient mutations detectable at short time frames, opposed to the fixed variation over longer periods of time. Moreover, not all regions of the genome present the same substitution rate, and thus hypervariable regions may add more diversity than the one expected from genome-wide mutation rates.

Higher variability may also be generated through other mechanisms, such as hypermutation or recombination. Hypermutation refers to the increase of mutation frequency due to changes in DNA mismatch repair systems, which can be a transient phenomenon or a stable and heritable mechanism (Jolivet-Gougeon et al. 2011). Recombination generates variability by replacing a fragment of the genome with a segment from a donor cell (intergenomic recombination) or from other regions within the genome of the same cell (intragenomic recombination). The exogenous transfer of genetic material from a donor cell to a recipient is referred to as horizontal gene transfer. Recombination and horizontal gene transfer have a big effect in generating variability as potentially large amounts of genetic changes are incorporated in a single event. In the presence of mixed infections, higher rates of variability can be generated by the horizontal gene transfer and recombination between the different strains present in the host, as it is been described in herpesvirus (Thiry et al. 2005), *Helicobacter pylori* (Kennemann et al. 2011), or human cytomegalovirus (HCMV) (Cudini et al. 2019).

The within-host pathogen subpopulations frequency dynamics are determined by genetic drift and natural selection (Didelot et al. 2016). Genetic drift describes the stochastic change in the frequency of genetic variants within the population, either towards fixation or disappearance. Genetic drift is considered to be an important force in shaping within-host diversity due to the common compartmentalization of pathogens in the host and fluctuations in population size. The effect of selection in genetic diversity is non-random and it depends on the host environment. Purifying selection reduces the amount of genetic diversity in the population by decreasing the frequency of those variants that have a deleterious effect. Similarly, positive selection increases the frequency of genetic variants that have a beneficial effect and increase the fitness of the pathogen in a specific environment. On the contrary, balancing and diversifying selection increase genetic diversity in the population. Balancing selection maintains multiple genetic polymorphisms in a population at frequencies larger than expected from genetic drift. Diversifying selection generates genetic diversity by recurrent positive or balancing selection. The role of selection is thought to be especially relevant in long chronic infections, especially in immunocompromised patients, where longer infection times promote the accumulation of genetic diversity or even the fixation of favorable variants such as the ones to avoid the immune system or to acquire antiviral resistance as observed in HIV and hepatitis C virus (Rambaut et al. 2004; Simmonds 2004) or antigenic selection in influenza virus (Xue et al. 2018).

Diversifying selection promotes the maintenance of multiple genotypes within a population. One example of this phenomenon is heteroresistance. Heteroresistance occurs when a pathogen population contains subpopulations with lower susceptibility to antimicrobials compared to the dominant population (El-Halfawy and Valvano 2015). Heteroresistance is often unstable, with the resistance population reverting back to susceptibility in the absence of antimicrobial in just a few generations (Nicoloff et al. 2019). This phenotypic instability in large populations facilitates adaptability

of the population as a whole in scenarios of changing environment. Heteroresistance has been identified as a relevant mechanisms in resistance to carbapenems (Andersson et al. 2019), as well as in infections of *H. pylori* (Mascellino et al. 2017), *M. tuberculosis* (Zheng et al. 2015) or *S. aureus* (Dordel et al. 2014). Diagnosis of heteroresistance is critical in clinical microbiology, as increasing evidence shows that the acquisition of drug resistance and treatment failure may be enhanced in patients with heteroresistance infections (Band and Weiss 2019).

The coexistence of multiple populations within a host and the evidence of diversifying selection as an important force shaping their evolutionary dynamics has prompted the development of within-host pathogen evolutionary models that incorporate the quasispecies concept. The quasispecies model represents the within-host pathogen population not as distinct genotypes, but as a distribution of genomes that acts as a unit of selection (Eigen and Schuster 1977; Domingo 1998). A consequence of this model is that genome sequences are indeterminate, and genomic variations reside within dynamic swarms of mutant pathogens (Domingo et al. 2012). Although the quasispecies model of evolution has been mostly proposed for rapidly evolving RNA viruses (Domingo et al. 1997), the recent discovery of the true extent of within-host diversity in bacterial pathogens (Didelot et al. 2016) may allow to revisit and apply the quasispecies model of evolution to a wider range of pathogens (Bertels et al. 2017). The quasispecies model allows to accommodate the rapidly evolving phenotypes that are observed in many pathogens, including antigenic selection, drug resistance, alterations in cell tropism, virulence or host range, and it will have consequences in many aspects of the clinical microbiology pipeline, from diagnostics and taxonomic classifications, to phenotype testing for virulence, antimicrobial resistance, or antigenic selection for vaccine candidates (Domingo 1998).

Pathogen genetic diversity can also be established within the host through the simultaneous co-transmission of an already diverse population from the donor (Fig. 1.3).

The amount of diversity passed from the transmission source to the recipient depends on the bottleneck size. The stringency of the transmission bottleneck determines how much of the pathogen diversity from the donor is transmitted and established in the recipient, and thus has a stochastic effect in the allele frequency and the evolutionary dynamics of the pathogen population along a transmission chain. Narrow, stringent bottlenecks limit the number of pathogen particles that are transmitted, and therefore have a higher potential to reduce the population diversity, affect the allele frequencies of some genetic variants and amplify the effect of genetic drift. Loose bottlenecks allow for a bigger population to be transmitted and have a lower impact in genetic diversity, and thus the population in the infection recipient should resemble more that of the donor at the time of transmission. A bottleneck is not exclusively limited to the time of transmission, but also within the recipient as many barriers, including physical defenses to colonization, immune system, competition, sanitation or medical intervention, may reduce the diversity of the newly transmitted population (Didelot et al. 2016). A consequence of this is that in some instances bottlenecks are not random, as some sub-populations can be genetically more adapted to evade the immune system or established infection, as shown in influenza virus (Xue et al. 2018). Bottleneck size is also influenced by transmission route. For instance, animal studies have proposed that transmission by direct contact results in a looser bottleneck than aerosol transmission in influenza (Varble et al. 2014; Frise et al. 2016). Moreover, environmental conditions may also influence bottleneck size. For instance, influenza transmission is favored in cold and dry conditions (Lowen et al. 2007), which in turn may increase the probability of more viral particles passing to the recipient. Given the multiple factors affecting bottleneck size, and its inherent stochasticity, modeling bottlenecks size is challenging, and estimates often vary.

An individual may also harbor a diverse pathogen population as a consequence of superinfection, which occurs when an individual is infected by multiple infections from different sources. Superinfection has been described as an important mechanism

in *S. aureus* (Mongkolrattanothai et al. 2011) or *M. tuberculosis* (Nardell et al. 1986).

Effects of within-host diversity in clinical microbiology

Diagnostic clinical microbiology is not likely to be affected by the presence of within-sample heterogeneity except in the case of mixed infections, when organisms from different species are present. One notable example is the detection of malaria parasites. Studies in multiple locations have shown that mixed *Plasmodium* species infections are frequent (Gupta et al. 2010; Sitali et al. 2015; Nundu et al. 2021), but often misdiagnosed and underestimated (Kotepui et al. 2020b). Microscopic diagnosis is currently the gold standard for malaria detection and diagnosis, but mixed infections of different *Plasmodium* species are often unrecognized. Lateral flow rapid tests are widely used for the detection of malaria, though their utility in detecting mixed infections may be into question given their consistent false positive rate for non-falciparum species in the presence of *P. falciparum* (Maltha et al. 2010; Berzosa et al. 2018) and their lower sensitivity at low parasite density. Accurate detection of *Plasmodium* mixed infections is critical to ensure appropriate treatment and malaria control measures. Mixed infections have been associated with severe malaria presentations and relapses among children (Genton et al. 2008; Kotepui et al. 2020a), and correct diagnosis is necessary to design an appropriate course of therapy (Kotepui et al. 2020b). Misdiagnosis of mixed *Plasmodium* infections can lead to drug resistance due to differential antimalarial efficacy in certain *Plasmodium* species (Gupta et al. 2010). Furthermore, malaria control measures and epidemiological monitoring can be affected if the prevalence of some *Plasmodium* species is underestimated.

The effects of within-host diversity in phenotype testing remain unclear. Clinically, perhaps the most concerning consequence would be the misdiagnosis of drug suscep-

tibility due to a mixture of populations with different susceptibility profiles, especially if drug resistant populations coexist at low frequency, which could lead to relapse with a drug resistant strain after antimicrobial treatment. In fact, reports of *M. tuberculosis* exogenous reinfections with drug-resistant strains after successful treatment of a drug susceptible infection could also be explained by the relapse of an undetected drug-resistant strain present at low frequency in the original infection, especially since coincidentally many exogenous reinfections are of drug- or multidrug-resistant tuberculosis (Small et al. 1993; Horn et al. 1994; Shafer et al. 1995; Turett et al. 1997; Niemann et al. 2000; Kaplan et al. 2003; Rie et al. 2005; Zong et al. 2018; Engelthaler et al. 2019). In principle, laboratory drug susceptibility testing should not be affected by mixed infections as the drug resistant population is expected to grow and increase in frequency in presence of antibiotics. However, if the population used for drug-susceptibility testing is not representative of the within-host population, laboratory methods could produce false negative results, especially since the isolates collected in the patient's sputum may not reflect the whole susceptibility profile of the *M. tuberculosis* subpopulations in the lung (Kaplan et al. 2003; Shamputa et al. 2006; Cohen et al. 2012). Molecular methods may not fare that well as they identify drug-resistance associated mutations which may not be detectable if present at low frequency in the mixed population. For instance, the *GeneXpert MTB/RIF* PCR based assay (Cepheid, Inc.) is a molecular rapid test that simultaneously detects the presence of *M. tuberculosis* and its susceptibility to rifampicin (Helb et al. 2010). Some reports have suggested that the sensitivity of the *GeneXpert MTB/RIF* assay to detect rifampicin resistance decreases in mixed *M. tuberculosis* infections (Zetola et al. 2014).

Within-host diversity has important implications for epidemiological and transmission studies. If multiple populations coexist within a host but only one is initially detected, a relapse with a different strain can be misdiagnosed as a new infection. Moreover, transmission events may be biased if infections are characterized by a single population (Worby et al. 2014). Typically, transmission and epidemiological studies use

a single genome for each isolate, representing the majority population within the clinical sample. The core assumption in transmission inference using genomic data is that sampled genomes within a transmission pair or an outbreak are genetically more similar than those that are not part of the transmission chain. Within-host diversity can thus hinder accurate reconstruction of transmission events if the population sampled from the transmission donor and the population transmitted to the recipient differ (Didelot et al. 2016). The extensive within-host diversity that is characteristic of chronic infections or prolonged carriage can further cause transmission events to be intractable. Sequencing multiple isolates per host can help characterize the full spectrum of variation and in turn reconstruct transmission chains with more precision. A methicillin-resistant *S. aureus* hospital outbreak investigation in the United Kingdom revealed the source of the outbreak by sequencing multiple colonies from a staff member (Harris et al. 2013). The staff colonies formed a cloud of diversity within which the patient's sequences clustered, revealing multiple transmission events and reintroductions.

Genetics and microbiology of *M. tuberculosis*

M. tuberculosis is the causative agent of tuberculosis, an airborne infectious disease estimated to cause 10 million new infections and 1.5 million deaths every year (World Health Organization 2021). Moreover, the increasing prevalence of antimicrobial resistance in tuberculosis poses an unsustainable human and economic burden and threatens the implementation of public health control measures.

The genus *Mycobacterium* comprises more than 170 species (Gagneux 2018). The species capable of causing tuberculosis disease are grouped into the *Mycobacterium tuberculosis* complex (MTBC). Human tuberculosis is mainly caused by *M. tuberculosis sensu stricto* and *Mycobacterium africanum*, though other animal-adapted members

of the MTBC can cause tuberculosis disease in humans, such as *M. bovis* (mostly infecting cows) and *M. caprae* (found in goats) (Shrestha et al. 2021). The human-adapted members of the MTBC are divided into eight phylogenetic lineages: L1 (Indo-Oceanic), L2 (East-Asian), L3 (East-African-Indian), L4 (Euro-American), L7, all of which belong to *M. tuberculosis sensu stricto*; lineages L5 and L6, referred to as *M. africanum* West Africa 1 and 2, respectively; and the recently described L8, which forms a sister lineage to the rest of the MTBC (Ngabonziza et al. 2020). Lineages have been further divided into sublineages based on insertion elements, deletions, variable number of repeats or the presence or absence of spacer oligonucleotides (spoligotyping), although single-nucleotide polymorphisms (SNPs) provide the highest resolution and robustness (Coll et al. 2014). Lineage 2 and lineage 4 within the MTBC are the most prevalent lineages, with lineage 2 dominating in East Asia and lineage 4 globally spread (Gagneux 2018).

For the last 80 years, a wide array of antibiotics effective against tuberculosis have been developed (Keshavjee and Farmer 2012). Currently, standard treatment for tuberculosis requires a combination of four drugs for a total duration of six months: Rifampicin (RIF, 6 months), Isoniazid (INH, 6 months), Pyrazinamide (PZA, 2 months), and Ethambutol (EMB, 2 months) (World Health Organization 2021). However, as these drugs were introduced into tuberculosis control programs, cases of drug resistance to each of them were being described (Gygli et al. 2017). Even with the widespread adoption of combinatorial drug therapy, multi-drug resistance (MDR) outbreaks have been reported. MDR tuberculosis is defined as a disease caused by strains of *M. tuberculosis* resistant to isoniazid and rifampicin—the two most effective first-line drugs—and it is estimated to affect 500,000 new patients annually (World Health Organization 2021). MDR tuberculosis is associated to worse clinical outcomes and higher toxicity as it requires longer treatments with second-line antibiotics. The emergence of extensively drug-resistant tuberculosis (XDR), defined as MDR in addition to any fluoroquinolone and at least one of three injectable second-line drugs

(amikacin, kanamycin, or capreomycin), further complicates patient management and tuberculosis control programs. Although the overall prevalence of TB is decreasing, the proportion of new cases that are rifampicin resistant or MDR is increasing with an estimated incidence of 450,000 cases (95% CI: 399,000–501,000) in 2021, while the number of detected DR cases is growing rapidly at a rate of 10-20% per year (World Health Organization 2021).

M. tuberculosis lacks horizontal gene transfer, it does not present detectable amounts of recombination and it does not have an accessory genome or vectors (Gygli et al. 2017). Drug resistance in MTB occurs almost exclusively by DNA mutations, such as SNPs and insertions and deletions (indels), by changing the drug target site, disrupting the prodrug-activating enzymes, overexpressing drug targets or by unspecific mechanisms such as overexpression of efflux pumps. Recent advances in bacterial whole-genome sequencing have resulted in a comprehensive catalog of mutations associated to antibiotic resistance (The CRyPTIC Consortium and 100000 Genomes Project 2018), which increases our understanding of drug resistance mechanisms and facilitates rapid drug susceptibility testing to support patient management and tuberculosis control programs.

Genomic epidemiology during the COVID-19 pandemic

Genome sequencing and genomic epidemiology played a critical role during the COVID-19 pandemic. The first genome sequence of SARS-CoV-2 was released to the scientific community in January 2020 (Wu et al. 2020), and by 2022 more than 10 million sequences were deposited on the GISAID database (Elbe and Buckland-Merrett 2017). The unprecedented availability of genomic data allowed researchers to track the global dispersal and evolutionary dynamics of SARS-CoV-2, while real-time genomic sequencing informed and influenced public health decisions during disease outbreaks (Attwood et al. 2022).

SARS-CoV-2 belongs to the viral family Coronaviridae, a family of enveloped, positive-strand RNA viruses that infect a wide range of hosts (Balloux et al. 2022). Early genomic data revealed that the genetic diversity of SARS-CoV-2 was initially low, accumulating mutations at a substitution rate of around 25–30 mutations per genome per year (Harvey et al. 2021). Despite its low mutation rate, phylodynamic analysis of SARS-CoV-2 genomes show some global transmission trends. For instance, genomic data discriminated between ongoing cryptic community transmission and independent introductions events (Fauver et al. 2020; Silva Filipe et al. 2021). The evolutionary dynamics of the different lineages also shed light on the effects of travel restrictions. Early lineages were highly cosmopolitan, while later lineages became more geographically structured with country- and continent-specific genotypes, reflecting the decline in international travel (Attwood et al. 2022). The coexistence of numerous, geographically restricted SARS-CoV-2 lineages came to an end when a series of lineages succesively and rapidly swept the world, rapidly increasing in frequency and replacing most lineages (Balloux et al. 2022). The term variants of concern (VoCs) was coined to described those lineages that rapidly increased in frequency. The first VoC was Alpha (initially designated B.1.1.7), which emerged in the UK and rapidly spread globally due to its higher transmissibility (Davies et al. 2021b; Volz et al. 2021b). Soon after, the Alpha VoC was replaced by the Delta VoC, first identified in India also in 2021 (Elliott et al. 2021). However, Delta VoC was also replaced by Omicron VoC and its sublineages, first described in South Africa in late 2021 (Viana et al. 2022).

Genomic data also allowed researchers to complement or even replace epidemiological data to study SARS-CoV-2 local transmission patterns. For instance, using genetic sequences, researchers were able to distinguish between hospital and community transmission in South Africa and The Netherlands (Sikkema et al. 2020; Giandhari et al. 2021). Similar findings allowed researchers to inform control programs using

genome sequencing data (Attwood et al. 2022).

COVID-19 was the first pandemic of the post-genomic era. The rapid deployment of genomic epidemiology has contributed to the understanding of the biology of SARS-CoV-2, its genetic diversity and evolution. Genetic data also catalyzed the development of SARS-CoV-2 vaccines, and its currently used to informed vaccine updates to cover new circulating variants. As the pandemic continues, whole-genome sequencing will remain as an essential tool to aid real-time surveillance and understand the long-term evolutionary dynamics of SARS-CoV-2.

Thesis statement

Whole-genome sequencing is becoming an essential part in the diagnosis and management of infectious diseases, with applications ranging from taxonomic classification of pathogens to transmission and contact tracing. In this thesis, I propose methodological and conceptual advancements for the use of whole-genome sequencing in clinical microbiology using public datasets as well as newly collected and sequenced data from two pathogens that pose a significant public health burden: *Mycobacterium tuberculosis*, the causative agent of tuberculosis; and SARS-CoV-2, the coronavirus that causes COVID-19.

First, I explore the effects of genetic diversity in commonly used reference mapping approaches using *M. tuberculosis* as an illustrative case (chapter 2). Moreover, I develop an approach to overcome the reference bias that also allows to analyze large number of isolates within comparative projects.

In chapter 3, I present an advancement in the diagnosis and theory of drug resistance by proposing and demonstrating the existence of pre-resistance mutations and genetic diversity that increases the risk of drug susceptible bacteria to acquire drug resistance

in the future. To do that, I describe a phylogenetic time-to-event analysis that predicts the evolution of drug resistance from drug susceptible ancestors, and apply it to show genetic markers of pre-resistance.

Finally, chapter 4 investigates the extent and transmission potential of within-host diversity, as well as its effects in genomic and phylogenetic inference. This chapter begins describing the extent of within-host diversity in SARS-CoV-2, as well as its stability and transmission. I then propose a model to include within-host diversity directly in phylogenetic and transmission inference.

I end with a general discussion of the work presented in this thesis and some considerations regarding the future direction of whole-genome sequencing in clinical microbiology.

Chapter 2

Reducing short-read mapping bias in *M. tuberculosis* using de-novo assembly

Role in the study: For this chapter, I conceptualised the study, designed the methodology, identified relevant publicly available datasets, analysed the data and interpreted the results under the guidance of the thesis supervisors.

Introduction

Whole-genome sequencing has become a critical tool in pathogen microbiology. The use of genetic data in *M. tuberculosis* is currently standard in many clinical and public health settings. Over the last decade, researchers have identified drug resistant associated variants that are used alongside culture methods for DST (The CRyPTIC Consortium and 100000 Genomes Project 2018) and, in some settings, have completely replaced them for the detection of resistance to first-line drugs (Meehan et al. 2019). Moreover, genomic sequences are now used in place of spoligotyping for genotyping studies and they are increasingly used for surveillance and contact investigation (Satta et al. 2018). As whole-genome sequencing is increasingly used to

influence patient outcome and public health practices, the robustness of the results becomes more crucial.

Most bioinformatic workflows in *M. tuberculosis* are based on resequencing approaches, which consist of genome assembly guided by a reference genome. This strategy is favored in *M. tuberculosis* due to the existence of a well annotated reference genome since 1998 based on the H37Rv strain (Cole et al. 1998), a low mutation rate and a high clonality characterized by undetectable recombination (Eldholm and Balloux 2016). However, some regions present higher variability than others (McEvoy et al. 2012), which may hinder correct mapping of the short reads to the reference genome. Moreover, duplicated regions also bias short-read mapping, as reads map to several locations with equal mapping quality. As a consequence, most bioinformatic workflows in *M. tuberculosis* remove problematic regions such as the *pe* and *ppe* gene families (Walter et al. 2020). The mapping bias is exacerbated by the fact that all lineages of *M. tuberculosis* are aligned to the same H37Rv reference, which belongs to lineage 4.9 (L4.9). Therefore, strains highly divergent to the reference genome are expected to be more inaccurate.

Using a common reference genome is nevertheless often necessary in *M. tuberculosis* research as many lineages do not have well annotated references. Moreover, whole-genome comparative projects require all sequences to be within the same genomic coordinates. For example, genome-wide association studies (GWAS) compare nucleotides at each polymorphic position in order to link genetic variation to phenotypes of interest such as drug resistance or pathogenicity. Similarly, having all samples within the same genomic coordinates is a prerequisite for phylogenetic inference since it studies the relatedness of genetic sequences by juxtaposing them base by base. Given that most genomic pipelines currently analyze whole-genome sequences of hundreds and thousands of isolates, alignment of all the sequences is computationally intensive, harder to parallelize and it is not realistic for aligning complete genomes.

Therefore, most projects involve mapping of short-reads to an assembled genome and inferring sequences in reference to it.

In this chapter, I investigate how the genetic diversity in *M. tuberculosis* affects the reconstruction of genome sequences using short-reads and characterize regions of high mapping errors in the *M. tuberculosis* genome by using simulations of short-reads from completely assembled genomes. Additionally, I propose that an approach using *de novo* assembly and mapping of short-reads to a sample specific reference genome reduces the number of errors in many of the problematic regions. The proposed method first performs *de novo* assembly of short-reads into longer contigs, which are in turned mapped to the H37Rv reference genome, taking advantage of the fact that longer sequences will increase the probability of mapping uniquely and with good quality to the target sequence. From the mapped reads, a sample specific reference genome is inferred, against which the short-reads are subsequently mapped. I then demonstrate than using a reference tailored to the isolate will allow for a better characterization of the *M. tuberculosis* genome and reduce mapping bias from short-read sequencing technologies.

Sample collection

A dataset of 337 *Mycobacterium tuberculosis* complete assemblies from 96 different projects was collected from the NCBI's Assembly portal database (Table A2.1). The majority of the assemblies were sequenced with third generation sequencing methods using the PacBio platform (123/337) and Oxford Nanopore methods (18/337). A total of 111 genomes were assembled using illumina short-read platforms. A combination of short- and long-reads was used in 63 samples.

Using lineage defining SNPs (Coll et al. 2014), it was inferred that the majority of the assemblies belonged to *M. tuberculosis* lineage 4 (48%, 161/337), while 10%

were lineage 1 (31/337), 27% were lineage 2 (92/337) and 4% were lineage 3 (14/337). The remaining assemblies belonged to minor lineages (4%, 12/337) as well as animal-adapted lineages such as *M. bovis* and *M. microti* (8%, 26/337).

To ensure high quality sequences, only assemblies sequenced with at least one of PacBio (Pacific Biosciences), Sanger sequencing, or a combination of Nanopore and illumina were kept, which resulted in 183 sequences included in the analysis. A total of 160 assemblies from *M. tuberculosis* lineages 1, 2, 3, and 4 were kept for further analysis.

Mapping reference bias in *M. tuberculosis*

To understand the intrinsic bias of mapping short-reads to the H37Rv reference genome, illumina short-reads were simulated from the initial assemblies, which permits comparing the results to a known starting genome. For each assembly, 10 sets of short-read sequencing raw reads were simulated using ART v2.5.8 (Huang et al. 2012) with the built-in quality and error profile for HiSeqX TruSeq Illumina sequencing systems. Mean read length of the simulated reads was set to 150bp, mean coverage of 50x, and mean and standard deviation for insert sizes between pairs of 650bp and 150bp, respectively. The simulated reads were mapped to the H37Rv genome using two different workflows (Fig. 2.1): a frequently used mapping approach where short-reads are directly mapped against the H37Rv reference genome; and an alternative workflow that adds a first step of *de novo* assembly to construct a sample-tailored reference genome, against which short-reads are mapped.

For the mapping pipeline, short raw reads were trimmed for low quality bases using Trimmomatic v0.39 (Bolger et al. 2014) with a minimum mean quality per base of 20 in a 4-base wide sliding window. The 5 leading and trailing bases of each read were removed, and reads with an average quality lower than 20 were discarded. The

resulting reads were mapped to the H37Rv reference genome (Gene bank: AL123456) using BWA v0.7.17 (Li and Durbin 2010) with the mem algorithm. The reads were processed using SAMtools v1.10 (Li et al. 2009), and duplicates were marked using Sambamba v0.8.0 (Tarasov et al. 2015). Genetic variants in reference to H37Rv were called using BCFtools v1.10.2 (Danecek et al. 2011), and variants were soft-filtered with a minimum mapping quality of 20, a base quality of 20, a minimum depth of 20, a minimum allele depth of 10 with 3 reads in both the forward and reverse strand, as well as position bias, mapping quality bias, strand bias.

The proposed alternative workflow extends on the mapping method by adding a first step of *de novo* assembly, where short-reads are assembled into longer contigs by finding overlaps between the reads. First, the short-reads were quality filtered using Trimmomatic v0.39 as described above. *De novo* assembly of the short-reads was performed using the SPAdes genome assembler v3.14.0 (Bankevich et al. 2012) using kmer sizes of 21, 33, 55, 77, and 101bp. The resulting contigs were filtered by removing contigs that are assembled with less than 2 reads and that have a length of 300bp. The filtered contigs were then mapped to the H37Rv reference genome using minimap2 (Li 2018) with the asm20. The *de novo* assembly step prior to mapping ensures a higher quality alignment, as the assembled contigs are longer than the original raw reads. Variants were identified using BCFtools as describe above, and a consensus sequence was inferred from the VCF file using Python custom scripts, with unmapped regions, gaps and ambiguous bases filled using the H37Rv reference. The inferred sequence was used as a sample-specific reference genome, against which the filtered short-reads were mapped using BWA and Sambamba and variants were called using BCFtools as previously described (Fig. 2.1). The quality of this last mapping is expected to be better as the sample raw reads are genetically closer to the sample-tailored reference genome than they are to H37Rv, increasing the probability of correct mapping.

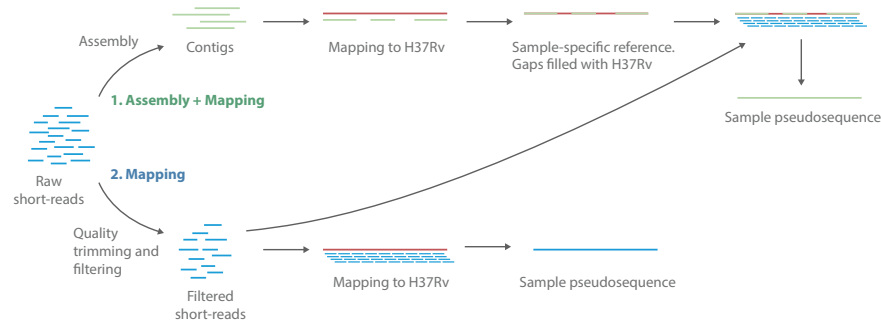


Figure 2.1: Overview of the two pipelines compared in this study.

Two different pipelines using illumina short-reads were compared. First, an assembly and mapping workflow, where short-reads are *de novo* assembled and then mapped against the H37Rv. Unmapped regions and gaps are filled using the H37Rv reference to construct a sample-specific reference genome against which short-reads are mapped. Alternatively, a common workflow of short-read mapping against the H37Rv reference genome.

The initial assemblies were mapped to the H37Rv genome using minimap2 to be used as the gold standard for comparison with the two workflows mentioned above. The VCF files generated from the two pipelines were compared to their respective initial assembly using the VCFR package (Knaus and Grünwald 2017) within the R statistical software. Structural variants, insertions and deletions were not considered for the comparison. Every position where either the assembly or the sample VCF showed an alternate allele were compared. Subsequently, the error rate was calculated for each gene and each sample. A false positive was considered when the sample showed an alternate allele when the assembly did not. A false negative was defined as the presence of a reference allele in the sample, while the assembly showed an alternate allele when compared to H37Rv. The error rate for each gene was defined as the number of false positives plus the number of false negatives, divided by the length of the gene.

To model the error distribution per gene and lineage, the error rate of all samples and

replicates was compared to a null distribution, which was estimated by calculating the error rate in 100 random genomic intervals of length equal to the specific gene. The fitted error rate and confidence intervals for each lineage were calculated by comparing the error rate to that of the null distribution in a linear model. Lineage 2 had the highest error rate (1.3×10^{-4} , 6.2×10^{-5} - 2.0×10^{-4} 95% CI), followed by lineage 4.8, lineage 4.4, and lineage 1 (all p-values ≤ 0.05 , Fig. 2.2a). Independent of lineage, higher error rates were estimated for *pe/ppe* gene families as well as insertion sequences and phages (Fig. 2.2b). Genes functionally related to cell wall processes had a higher error rate exclusively for lineage 2 samples (Fig. 2.2b).

Areas of the genome with high density of SNPs were correlated with a higher proportion of mapping errors. Using a genomic window of 10bp, the number of errors was inferred to rapidly increased as SNPs accumulate (Fig. 2.2c).

Hybrid assembly and mapping of short-reads decreases mapping errors

The error rate and the confidence interval of the error distribution using an assembly followed by mapping approach decreased for all lineages to a rate that was more uniform among them (Fig. 2.2a). Although error rates among *pe/ppe* and insertion sequences was still high when using the assembly and mapping approach, the error rate for cell wall related genes was reduced for lineage 2 to a level consistent with all other functional categories (Fig. 2.2a).

The impact of the different pipelines on phylogenetic inference was measured by inferring maximum-likelihood phylogenies using genome sequences from each pipeline and concatenating all sequences from both workflows. Phylogenetic trees were inferred using RAxML-NG (Kozlov et al. 2019) with a GTR+ γ , 20 starting trees (10 random and 10 parsimony), a minimum branch length of 10^{-9} and Stamatakis as-

approach and those inferred with the assembly and mapping pipeline were compared. Only 57% (90/159) of the sequences inferred from the same assembly shared a node in the phylogeny. The clustering bias was more pronounced in lineage 2, where the isolates tended to cluster by pipeline rather than by sample (Fig. 2.3b). Only 21% (15/72) of lineage 2 assemblies shared a node, while all samples from lineage 3 (10/10) and 91% (21/23) of lineage 1 clustered together. All isolates from sublineages 4.1, 4.4, 4.5, 4.7, and 4.8 were shared the same node, while 44% (4/9) of lineage 4.2, 69% (9/13) of lineage 4.3 and 86% (6/7) of lineage 4.9 were clustered together (Fig. 2.3c).

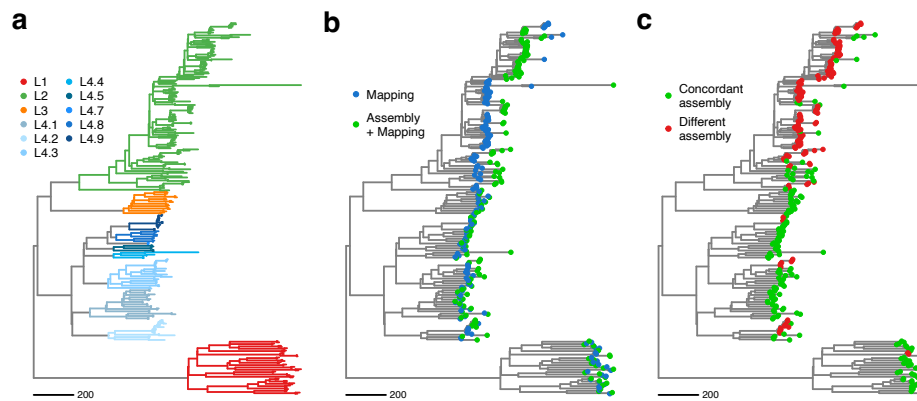


Figure 2.3: Phylogenetic inference of sequences derived from the mapping only workflow and the assembly and mapping approach.

Phylogenetic tree of 163 isolates, each of them with two sequences constructed with the mapping pipeline and the assembly and mapping one. Genetic distance is depicted in substitutions per genome. When there is little difference between the pipelines for an isolate, sequences derived from both methods are expected to cluster together. **a** Colors distinguish between *M. tuberculosis* lineages and sublineages. **b** Blue tips show sequences derived using the conventional mapping approach, while green tips represent those constructed with the assembly and mapping approach. **c** Isolates sharing a node between the different pipelines are shown in green, while isolates not clustered together are shown in red.

Discussion

The slow mutation rate and the low genetic diversity of *M. tuberculosis* contrasts with its functional and phenotypic diversity. Apart from drug resistance genetic variants, little is known about genetic determinants of many phenotypes of clinical and epidemiological interest, such as active disease, virulence or transmission. Given the limitations of short-reads in hypervariable or paralog regions, a high proportion of the *M. tuberculosis* genome is not well defined in many studies attempting to understand the genetic basis of pathogen phenotypes.

Most genomic projects in *M. tuberculosis* involve the study of multiple lineages. For comparative purposes, short-reads are aligned to the reference genome H37Rv irrespective of the lineage of the sample. Since H37Rv belongs to the sublineage 4.9 of *M. tuberculosis*, this choice of reference may affect alignment of short-reads when the query sample differs substantially from the target reference. In this chapter, I report lineage specific error rates when aligning short-reads to the reference genome H37Rv, and show that many of the alignment biases can be ameliorated when using a hybrid strategy of *de novo* assembly followed by short-read mapping.

The pipeline proposed in this chapter improves upon mapping of short-reads by including a step of *de novo* assembly. The resulting contigs are then mapped using the H37Rv genome as reference, filling in any gaps with H37Rv sequences, from which a sequence of the isolate within the H37Rv coordinates is derived. Finally, this sequence is used as a sample-specific reference genome to map the short-reads in order to correct the assembly, calculate quality metrics, and detect within-host diversity. The sample-specific reference genome increases the relatedness of the query and target sequences during short-read mapping and improves the characterization of highly divergent areas of the genome while maintaining H37Rv coordinates to allow for comparative analysis of multiple isolates. A hybrid assembly and mapping

approach has been proposed for the study of HIV genomes from short-read sequences (Wymant et al. 2018a). The work in this chapter focuses on the mapping bias in *M. tuberculosis* and proposes a similar pipeline using the H37Rv genome, which is the best annotated assembly and the most widely used reference in tuberculosis research.

The choice of reference genome affects the mapping of short-reads, especially as the distance between the query isolate and the reference increases (Walter et al. 2020). Using short-read simulations of well annotated assemblies, I show that lineage 2 presents a higher error rate when compared to lineage 4.9. Unsurprisingly, many of these errors accumulate in the *pe/ppe* gene families. The *pe/ppe* gene families represent about 10% of the *M. tuberculosis* genome (Cole et al. 1998) and are characterized by gene duplications and repetitive sequences, which make short-read mapping particularly challenging (Phelan et al. 2016). Similarly, I describe high error rates in insertion sequences and phages. Interestingly, genes linked to cell wall processes and host interaction had a high error rate exclusively in lineage 2 when compared to other lineages, especially genes from the *lpp* and *esx* gene families. Many of these errors are explained by the hypervariability of these regions and the accumulation of variants when compared to H37Rv. Most genome mappers currently used for whole-genome sequencing are local mappers that do not require the alignment of an entire read from the start to the end, removing sequences at the beginning and the end of the read (soft-masking). As SNPs accumulate in hypervariable regions, the penalties given by the mapper increase with each SNP, which can result in soft-masking of these regions by the mapper as excluding them gives a better alignment score than keeping the alignment with each SNP. Mapping short-reads to hypervariable regions or different lineages could be improved by changing the parameters and penalties of the aligner. But the ideal parameters may be different for each lineage, which could result in mapping irregularities and make isolates less comparable. Using the *de novo* assembly and mapping approach suggested in this chapter, lineage differences in error rates were reduced. Moreover, the error rate

in cell wall and host interaction related genes for lineage 2 was as low as for the remaining lineages. The results in this chapter show that mapping biases caused by hypervariable regions can be overcome using an additional step of *de novo* assembly in order to have larger query sequences and therefore increase the likelihood that they are uniquely mapped. Even though the mapping bias for *pe/ppe* gene families and phage sequences was also reduced with the recommended approach, their error rate remained high. Longer reads are thus required to better characterize these regions.

Phylogenetic inference was similarly affected by the reference genome bias. To study the effect of pipeline choice in the tree, alignments were concatenated for each assembly and pipeline and a phylogeny was inferred. If both pipelines were giving similar results, sequences from the same assembly would be expected to cluster together. This is indeed the case for a big part of lineage 1, lineage 3 and lineage 4. However, lineage 2 isolates clustered largely by pipeline, suggesting the genetic resolution gained by using the assembly and mapping pipeline was enough to guide sample clustering. Given that lineage 2 tends to display a lower genetic diversity when mapping short-reads to the H37Rv genome, changes in the sequence inference workflow would have a greater impact in clustering and phylogenetic methods.

Previous studies have shown that the choice of reference genome did not affect phylogenetic and epidemiological results (Lee and Behr 2016). However, this study was exclusively concerned on the effects on clustering of an outbreak of lineage 4 with relatively low genetic diversity. As all the isolates carry the same biases when using a specific reference genome, clustering may be consistent so phylogenetics analysis are not affected. In this chapter, I focused in high resolution comparative analysis in each position, which may be biased when using different reference genomes even if the broad clustering of the isolates is not. Even though clustering was sometimes affected by the choice of pipeline, the relationship of the isolates within each pipeline was consistent, which may explain why Lee and Behr (2016) did not find differences

in clustering when using different reference genomes. The results in this chapter are thus more consistent with recent publications looking at false negative and positive calls in short-read simulations (Walter et al. 2020).

Given that each dataset has different read lengths and read quality, using simulations allowed me to model the intrinsic error rate of genome mapping with a consistent read length and sequencing error rate distribution. However, other biases introduced during library amplification or sequencing, such as those in regions with high GC content, are not taken into account in the analysis, and therefore error rates in real data may vary from the estimates.

A focus on comparative genomic analysis meant that the initial assemblies had to be in the same coordinates as H37Rv so they could be compared with the sequences derived from the simulated reads. In order to do that, the complete assemblies were aligned against the H37Rv assembly and the resulting alignment was used as the gold standard. Even though the alignment is facilitated by the length of the query sequence, there may still be mapping errors that affect the comparisons.

These results highlight the inaccuracies of short-read mapping in some regions of the genome when using a unique reference genome, and how the approach proposed in this chapter decreases the error rate and reference bias. Decreasing the ambiguity in short-read sequences will allow researchers to include more regions for analysis, elucidating new phenotypic associations, aid transmission studies and increase our understanding of the mutational landscape of *M. tuberculosis* and other microorganisms.

Chapter 3

Genomic signatures of pre-resistance in *M. tuberculosis*

Role in the study: A version of this chapter was originally published in Torres Ortiz et al. (2021). As part of the study, I was responsible for the conception of the project, the design of the methodology, the analysis of clinical and whole-genome sequencing data, and the interpretation of the results.

Introduction

Mycobacterium tuberculosis is estimated to have killed 1 billion people over the last 200 years (Paulson 2013) and remains one of the world's most deadly pathogens (World Health Organization 2020). Drug resistance in bacteria imposes an unsustainable burden on health systems worldwide and compromises patient management and tuberculosis control programs. Recent advances in bacterial whole-genome sequencing have significantly improved the identification of drug resistance and have created a comprehensive catalog of antibiotic resistance genomic signatures, facilitating the implementation of rapid diagnostic tests in clinical microbiology laboratories (Walker et al. 2015). However, post hoc approaches to diagnosis miss the opportunity to preempt the emergence of drug resistance and implement preventive measures prior

to the acquisition and spread of antibiotic resistant disease.

An increased risk of drug resistance emergence is often attributed to inadequate implementation of control measures (Gandhi et al. 2010), but bacterial factors have also been proposed as potential contributors to drug resistance (Gygli et al. 2017). Evidence of differential drug resistance acquisition at the *M. tuberculosis* sublineage level is conflicting. Epidemiological and *in vitro* studies have suggested that the Beijing family, belonging to lineage 2, is hyper-mutable (Ford et al. 2013), with a propensity to develop resistance at a higher frequency than other lineages (Cox et al. 2005; Drobniewski 2005; Pang et al. 2012; Taype et al. 2012), while others cite evidence to the contrary (Glynn et al. 2002; Werngren and Hoffner 2003). Pre-existing resistance to one antibiotic (mono-resistance) is another factor that may influence the acquisition of multidrug-resistance (Borrell et al. 2013). Mono-resistance to isoniazid or rifampicin has been associated with increased rates of multidrug-resistance acquisition (Menzies et al. 2009; Saunders et al. 2011), but the relative risk of either remains unclear. Similarly, phylogenetic analyses suggest a stepwise progression towards multidrug-resistance, where mutations conferring isoniazid resistance tend to precede those linked to rifampicin resistance (Casali et al. 2014; Cohen et al. 2015; Eldholm et al. 2015; Manson et al. 2017).

Phylogenetic trees have been increasingly used to study pathogen dynamics and evolutionary processes of a wide range of phenotypes of epidemiological interest, including virulence and drug resistance acquisition (Galvani 2003; Grenfell et al. 2004). A necessary focus on improving the molecular diagnosis of drug resistance has led to the generation of large strain collections of drug resistant pathogens. However, unrepresentative samples of this kind enriched for drug-resistant isolates limit the ability to characterize the evolution and dynamics of drug resistance from a diverse background of ancestral susceptible strains. Inadequate sampling without comprehensive population level coverage or sufficient temporal span compounds this problem,

while the monomorphic nature of the *M. tuberculosis* genome makes constructing time-calibrated phylogenetic trees particularly challenging. As a consequence, a single mutation rate is often applied to the data, but this assumption inappropriately forces lineages and sub-lineages to conform to the same global mutation rate, thus limiting the inferences that can be made from the data.

In this chapter, I coin the term "pre-resistance" to encompass drug susceptible genotypes and mutations that have an increased risk to acquire antibiotic resistance in the future (Fig. 3.1). First, I hypothesise that some lineages in *M. tuberculosis* have a higher risk of drug resistance acquisition (Fig. 3.1a). Moreover, I also expand the term of pre-resistance to mono-resistance backgrounds, hypothesising that mono-resistance isolates have a higher risk of acquiring resistance to additional antibiotics than fully susceptible backgrounds (Fig. 3.1b). Finally, I apply the same concept to argue that some genetic polymorphisms in drug susceptible backgrounds can also lead to an increased rate of drug resistance emergence (Fig. 3.1c).

To understand the emergence and dynamics of pre-resistance, I present findings from samples collected over a 17-year time span with population level coverage in the hyperendemic suburbs of Lima, Peru. I apply a novel genome-wide survival analysis to a time-calibrated phylogeny of 3135 *M. tuberculosis* and show the existence of pre-resistance mutations among drug susceptible genotypes that increase the risk of future drug resistance emergence in *M. tuberculosis*. In this chapter, I demonstrate significant differences in the acquisition of drug resistance between lineages, on mono-resistant backgrounds, and at the level of nucleotide polymorphisms. The findings are then tested and replicated in an independent publicly available data set of 1000 whole genomes collected in Samara, Russia, and in a collection of 1573 isolates from different countries to demonstrate that they can be globally generalized.

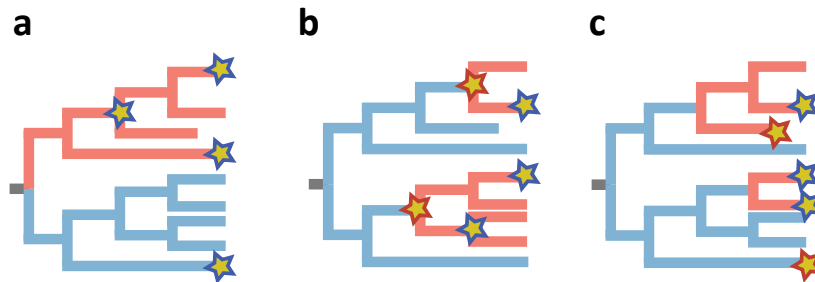


Figure 3.1: Phylogenetic analysis of pre-resistance

(a) Differences in drug resistance acquisition among lineages. Colors show two different lineages, and blue stars represent drug resistance acquisition events. The lineage marked as red has a higher rate of drug resistance acquisition than the blue lineage. (b) Mono-resistance to one drug (red star) increases the risk of acquisition of resistance to a different antibiotic (blue star). Red colors represent branches within the phylogenetic tree that have acquired mono-resistance, while blue branches are susceptible. (c) Genetic polymorphisms increase the risk of future drug resistance acquisition. Red color represents branches within the tree that are characterized by a specific mutation, while blue branches represent clades without that polymorphism. Red and blue stars mark drug resistance acquisition events.

Methods

Study design and sample selection

Samples were selected from previous projects taking place across the region of Lima. The first project consisted of a population level study carried out between 2008 and 2010 as part of the population level implementation of Microscopic Observation Drug Susceptibility (MODS) testing by Caviades et al. (2000) and Moore et al. (2006). A total of 2139 unselected patients of tuberculosis were collected. Of these patients, 284 were analyzed in previous studies (PRJEB5280) by Grandjean et al. (2017), while 1855 were processed as part of this project (PRJEB39837).

A second set of 213 MDR-TB samples was obtained from a 3-year long household follow-up study conducted between 2010–2013 by Grandjean et al. (2015), of which

185 randomly selected samples underwent whole-genome sequencing (PRJEB5280) (Grandjean et al. 2017).

Additionally, 42 unpublished whole-genome sequences of samples collected from 1999 to 2007 in different regions of Lima were added to the study (PRJEB47846), as well as 575 samples that were collected between 2003 and 2013 as part of the CRyPTIC Consortium (PRJEB32234) (The CRyPTIC Consortium and 100000 Genomes Project 2018), and 489 samples from the TANDEM Consortium (PRJEB23245) (Ugarte-Gil et al. 2020) taken between 2014 and 2016.

All samples without collection date, as well as reference clinical samples, were excluded from the final list. Drug Susceptibility Testing (DST) was performed either by MODS (Caviedes et al. 2000) or by the proportions method on agar (Clinical and Laboratory Standards Institute (CLSI) 2011).

To replicate the findings, all the analyses were repeated in a publicly available independent dataset from Samara, Russia (PRJEB2138), as the sampling was also representative of the population (Casali et al. 2014). A global data set of 1573 publicly available lineage 4 samples relatively enriched for drug resistance was also included as a validation dataset.

Whole-genome sequence analysis

Quality analysis of the raw reads was performed using FastQC (Andrews and Babraham Bioinformatics 2010). To ensure good quality alignments across genomic regions and between lineages, the methodology described in chapter 2 was used (Fig. 2.1). In short, a de novo assembly of the short-reads was done with SPAdes genome assembler v3.14.0 (Bankevich et al. 2012) across kmers of size 21, 33, 45, 55, 65, 75, 81, 101, 111, and 121. The resulting assembled contigs were mapped to the well annotated

H37Rv reference genome (Gene bank: AL123456) using minimap2 (Li 2018) with the asm20 option, while SNPs and indels were identified with BCFtools mpileup and BCFtools call v1.9 (Danecek et al. 2011) using the multiallelic calling algorithm, keeping the information about every single site in the genome in a VCF. Indels were left-aligned and normalized using BCFtools norm. A consensus sequence was created from the VCF file to be used as a sample tailored reference genome, against which the short-reads were mapped using the mem algorithm implemented in BWA v0.7.17 (Li and Durbin 2010). The alignments were sorted using SAMtools v1.9 (Li et al. 2009) and filtered for possible PCR and optical duplicates using Picard v2.19.0 (Thomer et al. 2016). Local realignment around indels was performed using the GATK v3.8-1-0 ‘IndelRealigner’ module (McKenna et al. 2010). The mean coverage for each sample was calculated as the number of mapped bases (excluding soft-clipped bases) divided by the genome size. Samples with a mean coverage lower than 15x were excluded from subsequent analysis. SNPs and indels were detected as described in the previous step.

Variants that did not meet the quality criteria were filtered using a combination of BCFtools filter and custom scripts in Python v3.7.3 with the following cutoffs: minimum Phred-scaled quality score (QUAL) of 20; minimum mapping quality (MQ) of 20; minimum genotype quality (QG) of 20; minimum read position bias (RPB), mapping quality bias (MQB), and strand bias (SP) of 0.001; minimum depth (DP) of 10 and a maximum of 5 times the mean coverage; minimum of reads supporting the alternate allele (AD) of 75% of the total depth in that position, with no less than two reads in the forward (ADF) and the reverse (ADR) strands. Additionally, SNPs within 2bp of an indel and indels within 3bp of another indel were removed, as both situations can be indicative of mapping artifacts. Positions that did not meet the quality criteria were annotated using the IUPAC ambiguity codes (Cornish-Bowden 1985). Samples with more than 25 high quality heterozygous calls were removed to avoid the inclusion of putative mixed infections. Variants that overlapped 100bp inter-

vals around known hypervariable regions, such repetitive elements and transposases (Cole et al. 1998), were removed from the analysis as this may affect the reliability of the alignment. Similarly, recombinant regions in genes coding for proline-glutamate (PE) or proline-proline-glutamate (PPE) (Phelan et al. 2016), and SNPs implicated in drug resistance (Coll et al. 2015) were excluded in order to minimize homoplasies that could disrupt the tree topology. The resulting sequences were concatenated to generate a multiple sequence alignment. Sites with a proportion of ambiguous bases higher than 10% were excluded from the analysis. Last, samples with a proportion of ambiguous sites in the alignment higher than 5% were excluded.

The functional consequence of variants was assessed using the Variant Effect predictor (VEP) v104.3 (McLaren et al. 2016)

Phylogenetic analyses

All the phylogenetic analyses were performed based on the alignment containing both L4 and L2 samples, as well as separately for each lineage.

A maximum likelihood phylogeny was inferred using RAxML-NG (Kozlov et al. 2019) with the GTR model, 20 starting trees (10 random and 10 parsimony), 100 bootstrap replicates, and a minimum branch length of 10^{-9} . An L2 sample randomly selected from the dataset was selected as an outgroup for the L4 phylogeny. Likewise, a random L4 isolate was used as a root for the L2 phylogeny. The tree containing both L4 and L2 samples was rooted using a lineage 1 isolate.

In order to construct a time-calibrated phylogeny, the existence of a detectable amount of evolutionary change between samples collected at different times was tested (Drummond et al. 2003; Biek et al. 2015). This was done in L4 and L2 separately in order to avoid population structure confounding in the temporal signal (Murray et al. 2016).

Two different tests of the temporal signal were applied (Fig. 1.2): the root-to-tip regression method and the date-randomization test (Rieux and Balloux 2016). For the former, BactDating (Didelot et al. 2018) was used to perform a linear regression between the collection dates and their root-to-tip genetic distance in the maximum likelihood tree. Additionally, a date-randomization test was performed, where evolutionary rates estimated by BactDating (Didelot et al. 2018) were compared between the observed data set and 100 data sets obtained by permutation of sampling dates (Duchêne et al. 2015).

BactDating (Didelot et al. 2018) was used to time-calibrate the tree using the mixed model for 10^7 iterations to achieve both convergence of the MCMC chains and an effective sample size of at least 100.

The phylogenetic global context of the Peruvian isolates was investigated by subsampling the phylogenies and repeating the analysis alongside publicly available isolates representative of the global diversity of *M. tuberculosis*. To subsample the phylogenies, first a random sample was selected from each phylogenetic cluster with a branch length lower than 1 SNPs per genome. The phylogeny was then divided into clusters of samples 50 SNPs apart, and a maximum of 20 samples for lineage 2 and 5 samples for lineage 4 were randomly selected for each cluster, unless it consisted of only 1 sample, in which case it was ignored. The phylogeny with the Peruvian subsamples and the global representatives was inferred separately for lineage 2 and for lineage 4 as described above.

The subsequent phylogenetic analysis was performed using the R package ape (Paradis et al. 2004). Marginal reconstruction of the ancestral sequences was carried out by maximum likelihood as implemented in Phangorn (Schliep 2011), including gaps and ambiguity codes to reflect prior probabilities of character states (Cornish-Bowden 1985).

Time-to-event analysis

Time-to-event analysis was performed on the tree using the R package *Survival* (Therneau and Grambsch 2000). The time was measured for all pairs of nodes as the distance between the older and the younger node in the time-calibrated phylogeny. An observation was defined as censored if both nodes were drug sensitive. On the other hand, an event occurred if the older node was drug sensitive and the younger node was drug resistant. Only the first acquisition of resistance was considered. Observations taking place before 1940 were discarded. The Kaplan-Meier survival curve and the Cox proportional hazard ratio were calculated. The Kaplan-Meier curves for different groups were compared using the log-rank test, where the null hypothesis is that there is no difference in survival between the different groups. Differences in the hazard ratio were tested using the likelihood ratio test. The entire pipeline was repeated for 100 phylogenetic bootstrap replicates.

Genome-wide association of pre-resistance

Missing base calls at the tips were imputed by maximum likelihood using the re-rooting method as implemented by Yang et al. (1995) and the IUPAC ambiguous codes to reflect tip state prior probabilities. In short, for each tip the phylogeny was re-rooted at that tip and the marginal probabilities for the missing bases were calculated for that node using the R package *phytools* (Revell 2012). Association analysis was performed at the gene and SNP level using the variant sites alignment for the tips of the phylogenetic tree, as well as the reconstructed sequences of the internal nodes. The phenotype was defined as leading to resistance in the phylogenetic tree, and thus only drug susceptible nodes that immediately preceded the first resistant node of each branch were considered. For the gene level analysis, non-synonymous variants were aggregated, excluding lineage specific SNPs. Loci with a frequency of non-synonymous variants in the dataset lower than 1% were not considered in the analysis. At the SNP-level, variants with a frequency in the population (tips of

the tree) lower than 1% were excluded. Furthermore, only those variants that were polymorphic at the node level were used. Genome-wide association was performed using the Cox proportional hazard model and the time between nodes. In order to correct for population structure, a genetic distance matrix was calculated using SNPs with a frequency in the population higher than 5%, and the eigenvectors were used as covariates in the Cox regression model. The genomic inflation factor (λ) was calculated as the ratio of the median of the empirically observed χ^2 to the median of the expected χ^2 . The p-values were corrected for multiple testing using a Bonferroni correction. Functional annotation of the genomic variants was assessed using Mycobrowser (Kapopoulou et al. 2011). Alignments were visually inspected for a random selection of samples using the Integrative genomics viewer (IGV) (Robinson et al. 2011) and the R software package *Gviz* (Hahne and Ivanek 2016).

Results

Population structure, genomic analysis, and patient demographics

A total of 3432 *M. tuberculosis* genomes from Lima (Peru) were analyzed, of which 3135 passed genomic quality filters. Of this, 2037 were part of a population level study carried out in 2009 where sputum samples were taken from all patients presenting tuberculosis symptoms in the Lima areas of Callao and Lima South (Grandjean et al. 2017). Comparison of drug resistance prevalence between the population level sampling using molecular genotyping and reports of epidemiological data in Peru published by World Health Organization (2020) are consistent: 1.5% (32/2037) of samples were rifampicin mono-resistant; 5% were isoniazid mono-resistant (105/2037), and 13% were multidrug-resistant (251/2037) (Table A3.1). The remaining samples were collected from cohort studies covering a 17-year period of research in the regions of Lima and Callao in order to achieve a sufficient temporal span in the sampling window (Fig. A3.1). Both lineage 2 and lineage 4 had a similar distribution

of sampling dates (Fig. A3.2).

The isolates were first aligned to the reference genome H37Rv using the approach described in chapter 2 (Fig. 2.1) to minimize mapping errors in L2 when using H37Rv as a reference genome. Lineages and sublineages were assigned using clade specific SNPs (Coll et al. 2014). Lineage 4 (Euro-American) consisted of 2807 samples, while lineage 2 (Beijing) had 327 isolates (Table 3.1). There was a single representative of lineage 1 (Indo-Oceanic), which was used to root the phylogenetic tree. The remaining samples from the data set, which included 5 *M. caprae* isolates, were not used in the downstream analysis. L4 had the highest diversity, comprising 1235 isolates from lineage 4.3 (LAM), 935 from lineage 4.1.2.1 (Haarlem), 271 from lineage 4.1.1 (X-Type), and 312 from lineages denoted as Type T, which encompasses lineages 4.5, 4.7, 4.8 and 4.9. Other minor sublineages included lineage 4.2.2 (TUR) and lineage 4.4. All L2 isolates were part of the Beijing sublineage (lineage 2.2), mainly from the sublineage 2.2.1 or Modern Beijing, and with only one representative of the sublineage 2.2.2 or Asia Ancestral (Ajawatanawong et al. 2019) (Table 3.1).

The alignment of the isolates to the reference genome resulted in 64,586 SNPs, of which 18,022 were singletons (28%). Most SNPs were not widely distributed across the population, and only 8088 variants had a frequency in the dataset higher than 1%. A total of 59,789 SNPs (16,934 singletons, 28%) were identified for L4, and 4821 SNPs (1370 singletons, 28%) for L2. The same analysis was applied to a publicly available data set from Samara, Russia, as a validation set where, unlike the Peruvian dataset, L2 constitutes the main lineage. A total of 28,414 SNPs were identified, consistent with previous publications of this data by Casali et al. (2014).

Patient demographic metadata was available for 2220 samples from the 2009 population level study, of which 88% were smear positive, 27% were previously treated for tuberculosis, and 2.8% were HIV positive, consistent with previous population

Table 3.1: Population structure.

	Clade name	Number
Lineage 1	Indo-Oceanic	1
Lineage 2	East-Asian	327
Lineage 2.2.1	Beijing	319
Lineage 2.2.1.1	Beijing	7
Lineage 2.2.2	Beijing	1
Lineage 4	Euro-American	2807
Lineage 4.1.1	X Type	271
Lineage 4.1.2.1	Haarlem	935
Lineage 4.2.2	TUR	3
Lineage 4.3.1	LAM	5
Lineage 4.3.2	LAM3	328
Lineage 4.3.3	LAM9	676
Lineage 4.3.4	LAM11	226
Lineage 4.4	-	51
Lineage 4.5	T Type	4
Lineage 4.7	T Type	31
Lineage 4.8	T Type	86
Lineage 4.9	T Type	12
Lineage 4 ¹	T Type	179

Lineages and sublineages defined using clade specific SNPs (Coll et al. 2014).

¹ Clade name assigned phylogenetically.

level estimates in Peru published by World Health Organization (2009). The median age was 28 years (IQR 21–41). In this cohort, 86% of L2 and 89% of L4 were smear positive.

Phylogenetic analysis and drug resistance emergence

The maximum likelihood phylogeny constructed using these alignments grouped the isolates by lineage similarly to previously published global data sets (Coll et al. 2014) (Fig. 3.2a, Fig. A3.3). To study the temporal dynamics of drug resistance acquisition at the population level, the ML phylogeny was time-calibrated using the sampling dates of the isolates, which extended from 1999 to 2016. Dated phylogenies were built separately for L2 and L4 in order to avoid the confounding effect of the temporal and population structures (Murray et al. 2016). Before time calibration of the phylogeny, the adequacy of the temporal correlation of evolutionary change was tested to reliably infer the model parameters. First, a root-to-tip linear regression (Fig. 1.2) of the number of substitutions from the root and the sampling times was fitted to confirm a positive association between time and evolutionary change. As uneven sampling may bias the root-to-tip regression (Rieux and Balloux 2016), a date-randomization test was additionally performed using the full bayesian model implemented in BactDating (Didelot et al. 2018) with the original dataset and 100 randomizations where the sampling times were permuted, representing the expectations of the model parameters in the absence of temporal signal (Fig. 1.2). The substitution rate estimated for the original dataset and for the 100 randomizations was compared to verify a lack of overlap between the 95% credible intervals. Both L4 and L2 datasets showed a clear temporal signal (Fig. A3.4), and thus model parameters could be confidently inferred from the data (Drummond et al. 2003; Biek et al. 2015). A relaxed clock model as implemented in BactDating (Didelot et al. 2018) was used, allowing the mutation rate to vary in each branch independently. The MCMC was ran until convergence of the chains was achieved, with an effective

sample size (ESS) of at least 100 (Fig. A3.5). The estimated rate for lineage 2 was 0.45 substitutions per genome per year (0.32;0.57 95% CI), while lineage 4 had a clock rate of 0.299 (0.25;0.36 95% CI) (Fig. 3.2b). The estimates of the molecular clock for both lineages were consistent with previous reports (Menardo et al. 2019). The most common recent ancestor (tMRCA) for the samples was placed at 565 CE (263;826 95% CI) for L4 (Fig. 3.2c), while L2 had a tMRCA in 1325 CE (1070;1499 95% CI) (Fig. 3.2d).

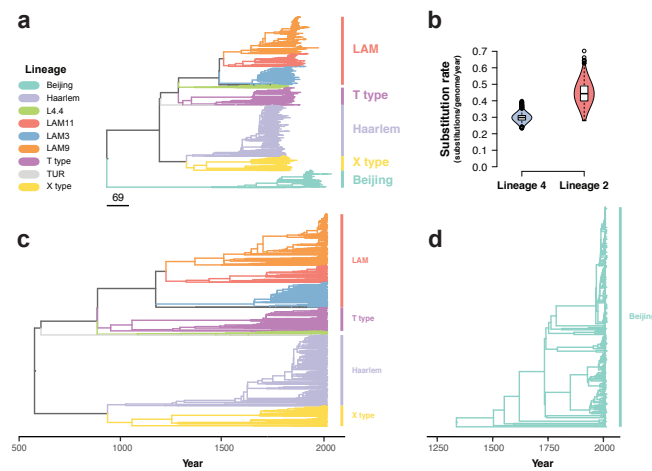


Figure 3.2: Phylogenetic analysis of 3134 *Mycobacterium tuberculosis* isolates from Lima, Peru

Colors represent different lineages and sublineages. (a) Maximum likelihood phylogeny. Scale in number of substitutions per genome. (b) Violin plots showing the posterior density distribution of the inferred substitution rate in substitutions per genome per year derived by sampling from 10^7 MCMC iterations. The substitution rate was estimated separately for lineage 4 (blue) and lineage 2 (red). Box plots inside the violin indicate the median value of the distribution (black horizontal line) and the interquartile range. Whiskers denote $1.5 \times$ the interquartile range, while outliers are plotted as individual points. (c) Time-calibrated phylogeny of lineage 4. (d) Time-calibrated phylogeny of lineage 2.

Drug resistance was inferred for all isolates at the tips of the phylogenetic tree using well-established drug resistance associated SNPs (Coll et al. 2015). In addition to

molecular typing of drug resistance, all isolates included in the analysis had DST performed either by MODS or by the proportional method in agar. DST and molecular typing showed consistent results for 96% of the samples for rifampicin resistance and 92% for isoniazid resistance. The time of emergence of drug resistant mutations was inferred by reconstructing the ancestral sequences of the internal nodes in the phylogenetic tree. The time of emergence of a specific antibiotic resistance mutation was approximated to the inferred year of the internal node where such mutation first appeared. The phylogenetic estimates of the emergence of drug resistance conferring mutations in lineage 4 occurred around the time of the known introductions of the corresponding drug. In contrast the emergence of drug resistance in lineage 2 was observed to have arisen many years after the introduction of antituberculous drugs. This is consistent with the geographic spread of lineage 4 in Europe together with early widespread use of drugs in this region (Table 3.2, Fig. 3.3). For both lineage 2 and lineage 4, the earliest inferred occurrence of resistance was to isoniazid, by the Ser315Thr mutation in the gene *katG*, around 1957 (1928;1978 95% CI) for lineage 2, and 1942 (1913;1960 95% CI) for lineage 4, in line with the reported wide introduction of isoniazid in 1952 (Keshavjee and Farmer 2012). The rifamycins were first isolated in 1957 (Keshavjee and Farmer 2012), and the first acquisition of resistance to rifampicin due to the *rpoB* mutation Ser450Leu was estimated to have occurred around 1951 (1931;1971 95% CI) for lineage 4 and 1974 (1953;1988 95% CI) for lineage 2. None of the drug resistant nodes reverted to susceptible along the branches of the two phylogenetic trees.

The emergence of compensatory mutations

It has been shown that secondary mutations arising after the acquisition of drug resistance may alleviate the fitness cost associated to antibiotic resistance mutations (Gagneux et al. 2006), but little is known about their temporal dynamics. Data on *M.*

Table 3.2: First emergence of drug resistance conferring mutations in Lima, Peru.

Drug	Lineage	Gene	Year	Mutation
RIF	Lineage 4	<i>rpoB</i>	1951.7 [1931.7 - 1970.7]	p.Ser450Leu
	Lineage 2	<i>rpoB</i>	1974.4 [1953.2 - 1986.8]	p.Ser450Leu
INH	Lineage 4	<i>katG</i>	1941.6 [1913.6 - 1959.7]	p.Ser315Thr
	Lineage 2	<i>katG</i>	1957.5 [1928.3 - 1977.4]	p.Ser315Thr
	Lineage 4	<i>embB</i>	1973.7 [1967.8 - 1980.8]	p.Gly406Ala
	Lineage 2	<i>embB</i>	1983.1 [1967.0 - 1994.1]	p.Met306Val
PZA	Lineage 4	<i>pncA</i>	1962.4 [1942.6 - 1972.2]	p.His51Arg
	Lineage 2	<i>pncA</i>	2002.1 [1997.0 - 2005.5]	c.-11A>C
STR	Lineage 4	<i>rpsL</i>	1974.5 [1955.9 - 1986.2]	p.Lys43Arg
	Lineage 2	<i>rpsL</i>	1958.1 [1934.6 - 1975.1]	p.Lys43Arg

Emergence of drug resistance conferring mutations for the 5 antibiotics historically used as first line drugs for the treatment of tuberculosis. RIF: rifampicin; INH: isoniazid; ETH: ethambutol; PZA: pyrazinamide; STR: streptomycin. Year presented as a point estimate with the highest posterior density interval.

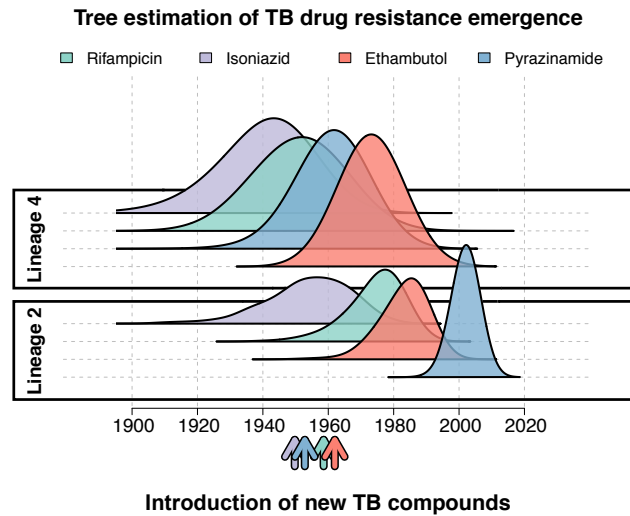


Figure 3.3: Inferred posterior density distribution of the earliest occurrence of resistance to first line antituberculous drugs

Posterior density distribution inferred using a time-calibrated phylogeny for both lineage 4 and lineage 2. Arrows represent the approximate time of antibiotic introduction.

tuberculosis drug resistance compensatory mechanisms is mainly limited to isoniazid and rifampicin resistance (Sherman et al. 1996; Comas et al. 2012). Non-synonymous mutations in the gene *rpoC* have been suggested as secondary compensatory mutations for rifampicin associated mutations in the *rpoB* gene (Comas et al. 2012). A total of 34% (258/755) of lineage 4 isolates harboring *rpoB* mutations also had *rpoC* non-synonymous polymorphisms; for lineage 2, 38% (33/87) of isolates with *rpoB* mutations carried *rpoC* polymorphisms. No significant differences were observed between lineages in a logistic regression model (OR = 0.85, 95% CI 0.54-1.35, p-value = 0.49). Overall, 62% of rifampicin resistant isolates carried Ser450Leu *rpoB* mutations (525/842). Rifampicin resistant isolates harboring Ser450Leu *rpoB* mutations had a higher probability of carrying mutations in the *rpoC* gene (52%, 272/525) than isolates with other *rpoB* mutations (6%, 19/317) in a logistic regression model (OR = 16.86, 95% CI 10.55-28.50, p-value = 4×10^{-29}). Only 3% (9/291) of the isolates

carried two non-synonymous mutations in the *rpoC* gene, while the rest had only one.

To understand the emergence of *rpoC* non-synonymous mutations, the phylogenetic branches of rifampicin resistant isolates were scanned from the root to the tip using the inferred sequences of the ancestral nodes to determine the time of emergence of *rpoC* non-synonymous mutations. The analysis was repeated in 100 bootstrap phylogenies to infer confidence values around the estimates. In both lineage 2 and lineage 4, the emergence of *rpoC* non-synonymous mutations occurred immediately after or at the same time as the emergence of rifampicin resistance, and continued steadily over time (Fig. 3.4). For lineage 2, there was not a single emergence of *rpoC* mutations occurring prior to the rifampicin resistance conferring mutations. In the case of lineage 4, two *rpoC* mutations emerged before rifampicin resistance: c.765150G>A and c.765590C>A. Both mutations appeared once independently in the entire phylogeny. The mutation c.765150G>A emerged in the dataset around the year 936 CE (727;1110 95% CI), in the MRCA of the clades X-type and Haarlem, all of which present the c.765150G>A mutation (1206/2807 isolates). For c.765590C>A, the estimated year of emergence was around 1867 CE (1813;1907 95% CI), and all the tips harboring the mutation belonged to a cluster of isolates part of the LAM11 sublineage (33/2807 isolates). Given that both mutations showed a clear phylogenetic structure and emerged independently only once in the dataset, they were considered phylogenetic mutations and were removed from the final analysis. The emergence of non-synonymous mutations in the *rpoC* gene was similar for isolates carrying the Ser450Leu *rpoB* mutation and for those isolates with other *rpoB* mutations (Fig. A3.6).

Phylogeographic history of *Mycobacterium tuberculosis* in Peru

To estimate the year of *M. tuberculosis* introductions to Peru, the Peruvian isolates were subsampled and analyzed alongside global representatives of both lineage 2

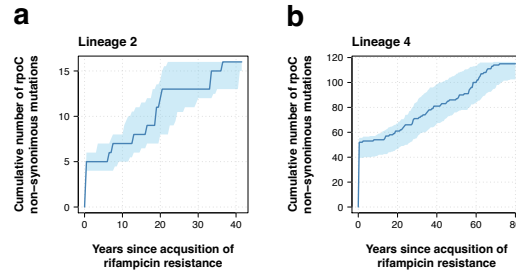


Figure 3.4: Dynamics of non-synonymous mutations in *rpoC*

(a-b) Cumulative number of non-synonymous mutations in *rpoC* over time. The x-axis represents the years since the inferred time of rifampicin resistance (time 0). Dark blue line shows the cumulative number of mutations for the ML tree, while the 95% confidence interval (shaded area) is inferred by repeating the analysis in 100 bootstrap phylogenies. The analysis was performed separately for lineage 2 **(a)** and lineage 4 **(b)**.

and lineage 4 for which both collection date and geographic origin were known. The phylogenies were time-calibrated using a relaxed clock model as implemented in BactDating (Didelot et al. 2018), and MCMC convergence was assessed using the traces of the model parameters (Fig. A3.7).

The phylogeographic history was inferred by reconstructing the ancestral states by maximum likelihood. The geographical origin of the isolates was treated as a discrete character, and it was assumed that the time of introduction occurred at the first Peruvian node of each clade.

Two early introductions of lineage 2 from China in 1880 (1856;1905 95% CI) and 1906 (1881;1929 95% CI) accounted for 84% of the Peruvian dataset (275/327 isolates), and 77% (90/117) of the drug resistance isolates as well as 77% (20/26) of the independent resistance evolutionary events (Fig. 3.5a). Later introductions to Peru from China occurred between 1981 (1969;1989 95% CI) and 2004 (1999;2007 95% CI), with one introduction in 1986 (1974;1994 95% CI) from South Asia representing 6 isolates, one of which is drug resistant.

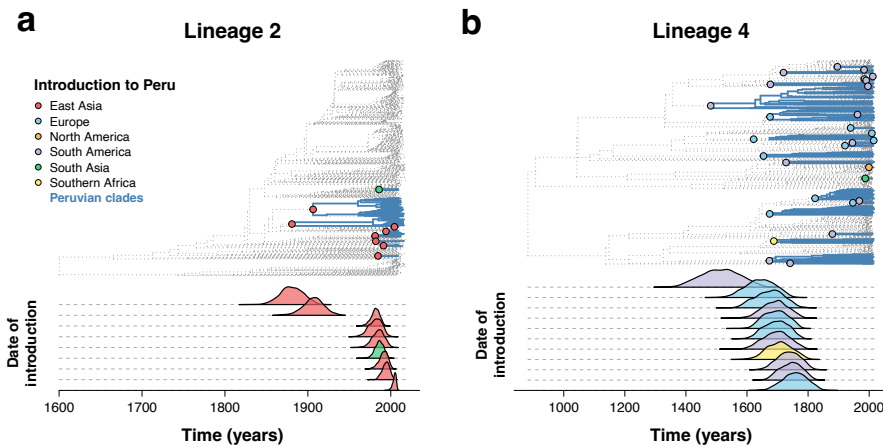


Figure 3.5: Phylogeographic analysis of *Mycobacterium tuberculosis* introductions to Peru

(a-b) Inferred introductions of *Mycobacterium tuberculosis* in Peru. The top part shows a time-calibrated phylogeny, with inferred introductions to Peru highlighted in the nodes with colors representing the country from which the clade was introduced. Peruvian clades are shown in blue. The bottom part shows the estimated year of introduction. The analysis was done separately for lineage 2 (a) and lineage 4 (b). For visual representation purposes, only the year of introduction of clades with more than 10 tips are shown for lineage 4.

Lineage 4 was inferred to have been introduced in Peru several times over the years, mainly from Europe and Brazil (Fig. 3.5b). Lineage 4.3 (LAM) represents the first and main introduction from Brazil around 1512 (1383;1598 95% CI), shortly after the well documented arrival of the Europeans to the continent. Subsequent smaller introductions from Europe and Brazil around 1644 (1545;1713 95% CI) and 1743 (1671;1796 95% CI) shaped the LAM and T type lineages. It was inferred that most of L4.1.2.1, part of the Haarlem sublineage, likely evolved from two introductions from Europe around 1693 (1614;1749 95% CI) and 1834 (1781;1869 95% CI). The most likely introductions of the X type clade (L4.1.1) occurred from Brazil in 1692 (1597;1754 95% CI) and South Africa for L4.1.1.3 in 1706 (1615;1773 95% CI).

Between lineage differences in drug resistance acquisition

The risk of acquiring drug resistance was calculated as the Cox Proportional Hazard Ratio (HR) using the time between sensitive internal nodes and the first drug resistant node in the time-calibrated phylogenetic trees. The Kaplan–Meier curves showed that lineage 2 had a higher probability of acquiring drug resistance than lineage 4 (Log-rank test p-value = 1.2×10^{-9} ; Fig. 3.6a). The estimated hazard ratio of drug resistance acquisition for lineage 2 was estimated to be 3.36 when compared to lineage 4 (HR 3.36, 95% CI 2.10 - 5.38, Likelihood ratio test p-value = 4.25×10^{-7}). A similar trend was observed in the Samara dataset (HR 4.82, 95% CI 3.74-6.21, Likelihood ratio test p-value = 6.8×10^{-34} ; Kaplan-Meier curve Log-rank test p-value = 1×10^{-39} ; Fig. 3.6b). The risk of drug resistance acquisition was also higher in L2 when compared to all sublineages of L4 in the Peruvian dataset, using LAM3 as a reference (L2 HR 3.32, 95% CI 1.84-6.28, Likelihood ratio test p-value = 1.9×10^{-4} , all other p-values > 0.2; Kaplan-Meier curve Log-rank test p-value = 6.9×10^{-8} ; Fig. 3.6c). To assess the adequacy of the data to the proportional hazard assumption, the relationship between the Schoenfeld residuals against time was calculated. In all cases, a non-significant association between the Schoenfeld residuals and time supported the use of the proportional hazards model (Fig. A3.8).

In order to evaluate the robustness of the maximum likelihood phylogeny, both the dating and the survival analysis were repeated on 100 phylogenetic bootstrap replicates. In both lineage 2 and lineage 4, the Cox Proportional Hazard ratio was not significantly different between the maximum likelihood phylogeny and the bootstrap replicates (Fig. A3.9a,b). Additionally, the Kaplan-Meier curve of the 100 replicates was similar to that of the maximum likelihood tree (Fig. A3.9c,d).

Both phylogenetic trees were subsampled to include only the isolates from the 2009 population level study to prevent any biases due to inclusion of datasets enriched for drug resistance isolates (Table A3.1). The results were congruent with those

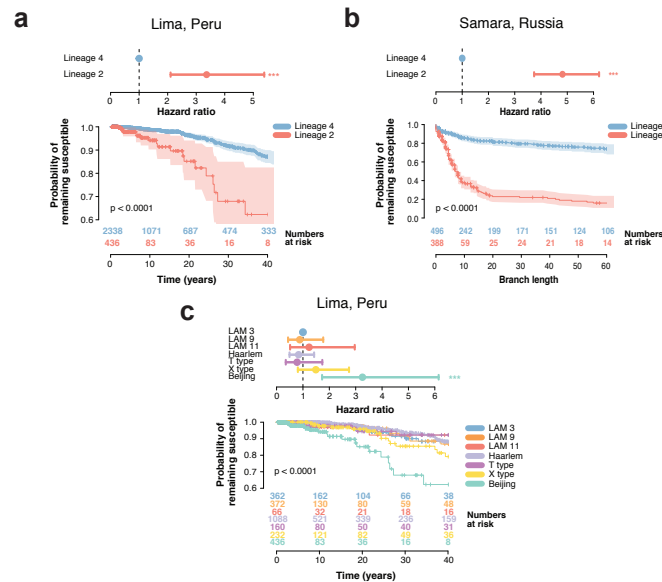


Figure 3.6: Hazard ratio and Kaplan-Meier curve for different sublineages of *Mycobacterium tuberculosis*

(a-c) Top: Hazard ratio (HR). Points and error bars represent the HR estimate and the 95% CI, respectively. The p-value for the HR was calculated using the likelihood ratio test. Bottom: Kaplan-Meier curve and numbers at risk. Y-axis represents the probability of remaining susceptible to any antibiotic, while the X-axis shows the time in years or the distance in branch length. Shaded areas show the 95% confidence interval. Kaplan-Meier curves were compared and p-values were derived using the log-rank test. (a) depicts the HR of lineage 2 compared to lineage 4 in the Peruvian dataset (HR 3.36, 95% CI 2.10 - 5.38, Likelihood ratio test p-value = 4.25×10^{-7}) and the different Kaplan-Meier curve for lineage 2 and lineage 4 (Log-rank test p-value = 1.2×10^{-9}). (b) same metrics for the Samara dataset (HR 4.82, 95% CI 3.74-6.21, Likelihood ratio test p-value = 6.8×10^{-34} ; Kaplan-Meier curve Log-rank test p-value = 1×10^{-39}). (c) shows HR between lineage 2 and the different sublineages of lineage 4 found in the Peruvian dataset (LAM9, LAM3, LAM11, Haarlem, X type and T type), using LAM3 as a reference (lineage 2 HR 3.32, 95% CI 1.84-6.28, Likelihood ratio test p-value = 1.9×10^{-4} , all other p-values > 0.2; Kaplan-Meier curve Log-rank test p-value = 6.9×10^{-8}). Statistical significance of the hazard ratio differences presented next to the CI bars (*, p<0.05; **, p<0.01; ***, p<0.001)

obtained with the entire dataset. Lineage 2 was characterized by a higher risk of drug resistance acquisition when compared to lineage 4 (HR 4.84, 95% CI 2.78 - 8.45, Likelihood ratio test p-value = 2.7×10^{-8} ; Kaplan-Meier curve Log-rank test p-value

= 7.9×10^{-10} ; Fig. A3.10a). Moreover, lineage 2 also had a higher hazard ratio than any sublineage of lineage 4 (lineage 2 HR 5.1, 95% CI 2.17-11.9, Likelihood ratio test p-value = 1.8×10^{-4} , all other p-values > 0.2; Kaplan-Meier curve Log-rank test p-value = 3.02×10^{-7} ; Fig. A3.10b).

Several confounders may also explain the differential rate of drug resistance acquisition between lineages (Fig. A3.11). There was no significant association between *M. tuberculosis* sublineages and HIV (Fig. A3.11a) and smear positivity (Fig. A3.11b) in a logistic regression model (n = 2133, all p-values > 0.1). It has been previously shown that prison conditions increase and amplify drug-resistance tuberculosis in Peru (Grandjean et al. 2015; Warren et al. 2018). This study population showed a higher proportion of prison infection in lineage 2 (5.8%, 12/206 patients) when compared to lineage 4 (1.4%, 28/2011 patients, Fig. A3.11c), and a higher risk of lineage 2 infection within the prison population (n = 2217, OR = 4.4, 95%CI 2.1-8.5, p < 0.001). This finding did not explain the higher rate of drug resistance acquisition in lineage 2, since all the lineage 2 samples taken from prisoners belonged to the same cluster and none of them harbored drug resistance conferring mutations. Since previous treatment with antituberculous drugs has been associated to an increased risk of acquiring drug resistance (Grandjean et al. 2015), the association between previous treatment and the different sublineages of this cohort was tested. Previous treatment history was available for 2236 samples contained metadata regarding previous treatment of tuberculosis. There was no significant association between *M. tuberculosis* sublineages and previous treatment with antituberculous drugs in a logistic regression model (n = 2236, all p-values > 0.1), suggesting that the between lineage differences in drug resistance acquisition observed in the survival analysis are not confounded by a differential distribution of antibiotics between sublineages (Fig. A3.11d). Behavioral patterns may also affect the dynamics of drug resistance acquisition and transmission. Patient sex (Fig. A3.11e) was not significantly associated with any of the *M. tuberculosis* sublineages (n = 2205, p-value = 0.3). The mean age

for patients with L4 infection was 33.3 (22-41 IQR), while the mean age for patients with L2 was 30.4 (21-36 IQR). Although the distribution of patient age was similar between the two lineages (Fig. A3.11f), lineage 4 showed an incident rate of age 1.09 higher than lineage 2 in a quasi-Poisson model (lineage 4 estimate = 0.09, standard error = 0.02, $p = 0.001$). On the other hand, age was not significantly associated with a higher risk of drug resistance acquisition in a logistic model (OR 0.999, 95%CI 0.994-1.003, $p = 0.69$).

The risk of developing MDR TB from isoniazid mono-resistance

To determine the effect of mono-resistance on the acquisition of further multidrug-resistance, the hazard ratio of acquiring rifampicin resistance was calculated for isoniazid mono-resistant ancestral genotypes versus susceptible ancestral strains. Genotypes with an isoniazid mono-resistant background had 15 times the hazard of developing rifampicin resistance tuberculosis relative to wild type susceptible strains (HR 15.12, 95% CI 10.54 - 21.69, Likelihood ratio test p -value $< 10^{-15}$; Kaplan-Meier curve Log-rank test p -value = 2.7×10^{-63} ; Fig. 3.7a). A larger hazard ratio was obtained in the Samara dataset (HR 37.28, 95% CI 18.81 - 73.88, Likelihood ratio test p -value = 3.4×10^{-25} ; Kaplan-Meier curve p -value = 4.6×10^{-63} ; Fig. 3.7b), although the low prevalence of mono-resistance clades in the Samara set may bias this estimate.

Multidrug-resistance was preceded by rifampicin mono-resistance only one time in the phylogenetic tree. Due to the infrequent occurrence of rifampicin mono-resistance prior to multidrug-resistance emergence, the risk of developing multidrug-resistance from a rifampicin mono-resistance background could not be reliably estimated.

Genomic signatures of drug resistance acquisition

Genome-wide survival analysis was performed using a Cox Proportional Hazard regression model to identify genetic variants in phylogenetic nodes inferred to be drug

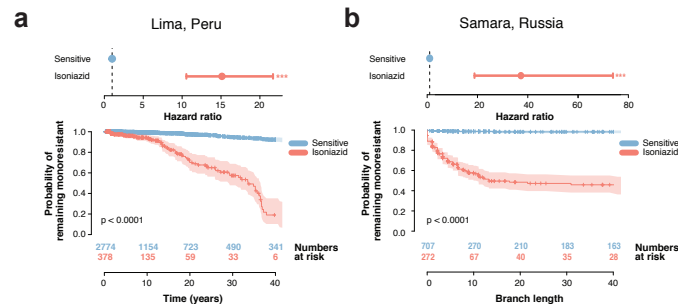


Figure 3.7: Hazard ratio and Kaplan-Meier curve for rifampicin acquisition

(a-b) Top: Hazard ratio (HR). Points and error bars represent the HR estimate and the 95% CI, respectively. The p-value for the HR was calculated using the likelihood ratio test. Bottom: Kaplan-Meier curve and numbers at risk. Y-axis represents the probability of remaining susceptible to rifampicin, while the X-axis shows the time in years or the distance in branch length. Shaded areas show the 95% confidence interval. P-values for the Kaplan-Meier curves were calculated using the log-rank test. (a) depicts the risk of acquiring rifampicin resistance from an already isoniazid mono-resistant background compared to a drug susceptible one (HR 15.12, 95% CI 10.54-21.69, Likelihood ratio test p-value = 1.3×10^{-40}) and the Kaplan-Meier curves for the different backgrounds (Log-rank test p-value = 2.7×10^{-63}). (b) same metrics for the Samara dataset (HR 37.28, 95% CI 18.81 - 73.88, Likelihood ratio test p-value = 3.4×10^{-25} ; Kaplan-Meier curve p-value = 4.6×10^{-63}). Statistical significance of the hazard ratio differences presented next to the CI bars (*, p<0.05; **, p<0.01; ***, p<0.001)

susceptible but associated with a higher risk of progression towards drug resistance. Resistant nodes were defined as those inferred to be resistant to any antibiotic in order to identify common pathways of increased risk of acquiring resistance regardless of the specific antibiotic. The association analysis was performed in lineage 4 and lineage 2 separately, and population structure was corrected further using a kinship matrix, which reduced the genomic inflation factor (λ) to 1.16 (Fig. A3.12).

Six variants in drug susceptible ancestral genotypes were associated with a higher risk of acquiring drug resistance in lineage 4 after population and multiple testing correction, three of which were in previously annotated genes (Fig. 3.8a). The variant with the lowest p-value corresponded to a 9bp deletion at location 2,604,157 in the locus *lppP*, which encodes a lipoprotein and has been predicted to be required for

growth in macrophages (Rengarajan et al. 2005). This deletion had a frequency of 1.7% in the population, and it evolved 12 times independently along the phylogenetic tree. Genotypes with this variant had a hazard ratio 7.36 times greater than those with an intact *lppP* (Fig. 3.8b, HR 7.36 95% CI 3.85-14.04, p-value = 7.46×10^{-10}). The findings were replicated in a global data set of 1573 L4 isolates, which was relatively enriched for drug resistance (55% of samples were resistant to any drug). The *lppP* deletion had a frequency of 9% and inferred susceptible genotypes with the deletion had a hazard ratio 3.6 times greater than those without it (HR 3.6, 95% CI 1.9-6.9, p-value = 8.7×10^{-5}). Two synonymous polymorphisms at *esx* genes were found to be associated with a higher risk of acquiring drug resistance in inferred drug susceptible genotypes. The *esx* gene family encodes protein secretion systems described to be critical for growth, pathogenesis, and mycobacterial–host interactions (Tufariello et al. 2016). The two polymorphisms were detected in the gene *esxL* at position 1,341,044 with a hazard ratio of 3.2 (HR 3.2 95% CI 1.91-5.37, p-value = 1.01×10^{-6}), and at position 2,626,011 in the gene *esxO* (HR 11.12, 95% CI 5.50-22.5 p-value = 1.52×10^{-5}) with a frequency in the population of 17% and 5%, respectively. In the L4 global dataset, the *esxL* SNP had a frequency of 19.5% while the *esxO* polymorphism was present in 10% of the isolates. Inferred susceptible genotypes in the global data set carrying the mutation in *esxO* had a risk of acquiring drug resistance 3.1 times higher than those with the reference genotype (HR 3.1, 95% CI 1.3-7.3, p-value = 0.009), while those carrying the mutation in *esxL* had a risk 1.4 higher, although differences were not statistically significant (HR 1.4, 95% CI 0.7-3.0, p-value = 0.3). Visual inspection of the short-read alignments around the described genes was undertaken to confirm high quality alignments over these regions (Fig. A3.13).

For the gene-based GWAS, non-synonymous variants were aggregated for each loci and a binary matrix was created reflecting whether internal nodes and tips contained at least one non-synonymous polymorphism for each gene. After population and

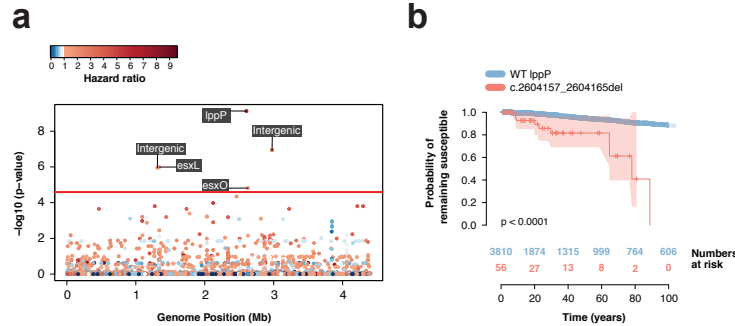


Figure 3.8: Genome-Wide association study (GWAS) results. (a) Manhattan plot for GWAS on increased risk of drug resistance acquisition in lineage 4. The red line represents the Bonferroni corrected p-value threshold of 3.37×10^{-5} . Labels show the genes where the significant hits are located. Colors indicate the hazard ratio, with a scale of blue representing hazards ratio lower than 1 and a scale of reds for hazard ratios higher than 1. (b) Kaplan-Meier curve and numbers at risk of a 9bp deletion in the gene *lppP* comparing the probability of remaining susceptible between those nodes without the deletion (blue) and those with it (red). Shaded areas represent the 95% confidence interval. The p-value for the Kaplan-Meier curves was calculated using the log-rank test.

Bonferroni multiple testing correction, a total of 35 variants had a p-value lower than the significance threshold of 4.38×10^{-5} (Table 3.3). Functional annotations curated in the Mycobrowser (Kapopoulou et al. 2011) showed that most genes were related to metabolism and cellular respiration (*frdB*, *cyp135A1*, *Rv3113*, *hemE*, *pyrG*, *galK*, *icd1*, *Rv1096*, *fum*, *galE2*, *Rv1751*, *Rv3712*, *asnB*) as well as cell wall processes (*kdpA*, *lppP*, *cfp2*, *lytB1*, *mmpL1*, *Rv1417*, *lgt*, *Rv0226c*, *caeA*, *murF*, *pstB*).

No significant associations were identified for lineage 2 after correcting for population structure, possibly due to the lower diversity of lineage 2 as well as the strong lineage effect on the phenotype. The analysis could not be replicated in the Samara dataset as the Samara dataset was significantly smaller and hence lacked sufficient statistical power.

Table 3.3: Gene-based association analysis.

Gene Name	Rv Number	HR	SE	P value ¹	Frequency	Functional category ²
Rv2510c	Rv2510c	16.74	0.4	1.79×10^{-14}	0.01	Unknown
yrbE2A	Rv0587	4.77	0.39	2.43×10^{-12}	0.01	Virulence, detoxification, adaptation
kdpA	Rv1029	1.46	0.81	3.56×10^{-10}	0.42	Cell wall and cell processes
lppP	Rv2330c	7.14	0.45	4.44×10^{-10}	0.02	Cell wall and cell processes
cfp2	Rv2376c	1.97	1.04	8.4×10^{-09}	0.03	Cell wall and cell processes
lytB1	Rv3382c	5.2	0.45	3.05×10^{-08}	0.01	Cell wall and cell processes
frdB	Rv1553	1.62	0.64	3.54×10^{-08}	0.42	Metabolism and respiration
mmpL1	Rv0402c	1.6	0.38	1.07×10^{-07}	0.06	Cell wall and cell processes
fadD5	Rv0166	2.27	0.47	1.57×10^{-07}	0.05	Lipid metabolism
cyp135A1	Rv0327c	2.07	0.4	2.03×10^{-07}	0.04	Metabolism and respiration
Rv1897c	Rv1897c	1.18	0.72	2.55×10^{-07}	0.16	Unknown
Rv3113	Rv3113	1.69	0.72	2.55×10^{-07}	0.52	Metabolism and respiration
gpdA2	Rv2982c	0.9	0.69	2.99×10^{-07}	0.05	Lipid metabolism
Rv0579	Rv0579	1.33	0.53	6.7×10^{-07}	0.02	Unknown
Rv1417	Rv1417	0.83	0.72	9.51×10^{-07}	0.16	Cell wall and cell processes
recD	Rv0629c	2.63	0.37	1×10^{-06}	0.03	Information pathways
Rv3903c	Rv3903c	1.14	0.28	1.05×10^{-06}	0.19	Unknown
hemE	Rv2678c	1.28	0.55	2.18×10^{-06}	0.02	Metabolism and respiration
lgt	Rv1614	0.83	0.64	2.25×10^{-06}	0.17	Cell wall and cell processes
pyrG	Rv1699	4.1	0.32	3.15×10^{-06}	0.04	Metabolism and respiration
galK	Rv0620	0.87	0.72	4.73×10^{-06}	0.16	Metabolism and respiration
Rv2915c	Rv2915c	4.8	0.36	7.24×10^{-06}	0.02	Unknown
Rv0226c	Rv0226c	2.87	0.41	7.73×10^{-06}	0.03	Cell wall and cell processes
Rv0021c	Rv0021c	1.04	0.62	8.6×10^{-06}	0.06	Unknown
caeA	Rv2224c	0.92	1.05	8.69×10^{-06}	0.02	Cell wall and cell processes
Rv1501	Rv1501	2.26	0.53	1.09×10^{-05}	0.08	Unknown
icd1	Rv3339c	4.42	0.53	1.14×10^{-05}	0.04	Metabolism and respiration
Rv1096	Rv1096	3.05	0.64	1.38×10^{-05}	0.02	Metabolism and respiration
fum	Rv1098c	0.9	0.48	1.56×10^{-05}	0.05	Metabolism and respiration
galE2	Rv0501	2.84	0.51	1.66×10^{-05}	0.01	Metabolism and respiration
Rv1751	Rv1751	4.43	0.42	1.67×10^{-05}	0.02	Metabolism and respiration
murF	Rv2157c	2.38	0.54	2.78×10^{-05}	0.03	Cell wall and cell processes
Rv3712	Rv3712	0.93	0.58	3×10^{-05}	0.05	Metabolism and respiration
asnB	Rv2201	3.62	0.48	3.14×10^{-05}	0.02	Metabolism and respiration
pstB	Rv0933	1.49	0.52	4.19×10^{-05}	0.07	Cell wall and cell processes

¹ P-values derived from the Likelihood ratio test.

² Annotations extracted from the Mycobrowser (Kapopoulou et al. 2011)

Discussion

The results presented in this chapter represent conceptual and methodological advancements in phenotype and drug resistance testing in clinical microbiology. Using *M. tuberculosis* outbreaks in Lima, Peru, I provide a description and an evaluation of pathogen pre-resistance (pre-existing polymorphisms that predispose to the acquisition of future drug resistance), which could facilitate preemptive actions in the control and treatment of drug resistance.

This study represents the largest population level genomic analysis of *M. tuberculosis* to date. The 17-year sampling time frame provided a unique opportunity to study drug resistance acquisition dynamics and evolution. Using a novel ancestral state genome-wide survival analysis to move in time through the phylogenetic tree, it is shown that *M. tuberculosis* is predisposed to acquire drug resistance mutations at the lineage level, after mono-resistance, and at the level of nucleotide polymorphisms. Understanding pathogen genetic factors that predispose strains to evolve drug resistance could help prevent the acquisition and spread of resistance by closely monitoring patients at greatest risk of developing resistance. Moreover, these genetic markers could prevent treatment failure by employing broader treatment regimens for those susceptible pathogens more likely to acquire drug resistance.

Previous studies of acquired drug resistance at the sublineage level in *M. tuberculosis* have led to contradictory outcomes, with small sample sizes in fluctuation assays (Werngren and Hoffner 2003; Ford et al. 2013) or using amalgamated sub-population level samples. The results presented in this chapter demonstrate that lineage 2 acquired resistance to antibiotics more rapidly than lineage 4. There were no significant differences observed in drug resistance acquisition between the sub-lineages of the most diverse lineage 4. Even though lineage 2 showed an increased risk in drug-resistance acquisition, lineage 4 evolved resistance earlier than lineage 2 for almost

all drugs analyzed, with the exception of streptomycin. This may be explained by the Euro-American distribution of lineage 4 and the earlier widespread implementation of antibiotics in these regions. The analysis also suggests that the acquisition dynamic of compensatory mutations was similar for both lineage 2 and lineage 4. After rifampicin resistance associated mutations evolve in a clade, non-synonymous mutations in *rpoC* start to occur and steadily accumulate over time. Thus, pre-resistance mutations emerge independently of compensatory mutations.

The identification and control of mono-resistant strains is a key component of tuberculosis public health infection prevention and control efforts. Mono-resistance is associated with worse clinical outcomes (Heijden et al. 2017) and an increased probability of progressing to multidrug-resistance (Saunders et al. 2011). At the population level, this study quantifies this risk and shows that isoniazid mono-resistant strains have at least 15 times the hazard of developing multidrug-resistance relative to wild type susceptible strains. Despite the use of new therapeutics, multi-drug resistant tuberculosis continues to require polypharmacy with increased toxicity and longer treatment duration (World Health Organization 2020). Globally, molecular rapid drug resistance surveillance is focused primarily on rifampicin, with the widely implemented GeneXpert MTB/RIF PCR based assay unable to detect isoniazid mono-resistance. Although Drug Susceptibility Testing (DST) is the current gold standard for identification of drug resistance isolates and can detect isoniazid mono-resistance, known diagnostic delays associated with it may limit its use in reducing mono-resistance amplification (Ahmad and Mokaddas 2014; Nguyen et al. 2019). Inadequate diagnosis of isoniazid mono-resistance will inevitably lead to inappropriate treatment and could fuel rapid evolution of multidrug-resistance, thus posing a significant threat to tuberculosis control.

In this chapter, I also identified loci associated with higher risk of future drug resistance acquisition. To remain polymorphic, these variants must be under balancing

selection and only be positively selected once exposed to drug therapy. Rather than causing resistance directly, these variants could promote resistance acquisition by compensating for the fitness costs of resistance *in vivo* (Gagneux et al. 2006) or by increasing drug tolerance (Hicks et al. 2018).

At the gene level, most non-synonymous mutations associated with pre-resistance genotypes were located in genes related to cell wall processes and metabolism. Functional studies and prospective clinical trials are warranted to confirm their association with future drug resistance acquisition.

The variant with the lowest p-value corresponded to a 9bp deletion in the gene *lppP* present at a frequency of 1.7% in the population and arose 12 times independently. Deletions in lipoproteins have been well characterized in the past (Tsolaki et al. 2004), and *lppP* has been predicted to be required for growth in macrophages (Rengarajan et al. 2005). Lipoproteins can act as antigenic proteins (Tsolaki et al. 2004), and thus deletions in the genes encoding them may alter the interaction between the bacilli and the macrophages, potentially conferring a selective advantage in the presence of drug and increasing the probability of acquiring drug resistance. This variant was also present in a global dataset for L4 at a frequency of 9%, which could be explained by the higher prevalence of drug resistance isolates in most publicly available data sets.

Two additional synonymous variants were identified in the genes *esxL* and *esxO*, which encode the ESAT-6-like proteins EsxL and EsxO. These genes are part of a family of genes that encode immunogenic secreted proteins that play a role in mycobacterial growth, pathogenesis, and host-pathogen interactions (Tufariello et al. 2016). Moreover, the EsxO protein has been associated with pathogenesis by inducing autophagy in infected macrophages (Mohanty et al. 2016). Synonymous homoplasic variants in *esx* genes have been previously identified (Grandjean et al. 2017), but

their phenotypic effects are still unclear.

This study benefited from an unbiased population level coverage of both drug resistant and drug susceptible strains that enabled us to reliably correct for the founder effect and control for the influence of pre-existing population diversity. The large sampling size and time frame — a consequence of 17-years of continued research in the same location — allowed us to generate time-calibrated phylogenies without imposing a global mutation rate. This was a pre-requisite for the downstream analyses and the novel GWAS survival analysis approach. I was also able to replicate the sublineage and mono-resistance dependent hazards of acquired resistance in the smaller Samara dataset. However, the time scale and size of this publicly available data was insufficient to allow us to confirm the effect of the *lppP* deletion in a second independent dataset.

Although this phylogenetic analysis reveals trends of drug resistance acquisition over evolutionary time, prospective cohort studies are required to determine the effect of these mutations at the individual patient and household level. Non-bacterial factors are unlikely to influence the findings, since they would have to have been disproportionately and consistently associated to a specific lineage over long periods of time. Nevertheless, the influence of confounding variables on the dataset was explored. There was no difference observed in the proportion of patients receiving previous antituberculous treatment between the two lineages. This makes the findings unlikely to be influenced by differential exposure to drugs among lineages. There was no difference in sputum smear grade between lineages, suggesting that the findings are not a consequence of increased pathogenicity of lineage 2 in comparison to lineage 4. Moreover, other factors that may affect the rates of drug resistance acquisition such as HIV status, sex, or imprisonment, did not show differences between lineages. Even though the age of patients with lineage 4 infection was significantly higher than that of patients with lineage 2, the difference was small. Additionally, patient age was not

associated to a higher incidence of drug resistance, and therefore it is unlikely that age differences explain the higher risk of acquiring antibiotic resistance of lineage 2. Differential healthcare systems influence the acquisition and transmission of drug resistance tuberculosis, and thus importation events to Peru of resistant strains from specific lineages could have affected the dynamics of drug resistance acquisition. Lineage 2 in Peru is characterized by two importations around 1900 CE, which is consistent with major immigration events of laborers from China to Peru at the end of the 19th century to work in the railroads, guano mines, and cotton and sugarcane plantations (Gonzales 1989; Ritacco et al. 2008). Conversely, the majority of lineage 4 clades were imported from Europe and Brazil between the 16th and 18th centuries, compatible with European colonial expansion (Brynildsrud et al. 2018). Therefore, significant immigration to Peru occurred well before the advent of antibiotics, which limits the influence of imported drug resistant strains. Moreover, the majority of introductions occurring in recent times were of drug susceptible clades. Nevertheless, it is possible that some resistance events may have arisen as a result of importation of resistant strains from countries with different drug selection pressures.

In summary, this population wide 17 year-long epidemiological study of *M. tuberculosis* genetics provides the first description and evaluation of pre-resistant polymorphisms in susceptible genotypes that predispose to the acquisition of future drug resistance. Prediction of future drug resistance in susceptible pathogens together with targeted expanded therapy has the potential to prevent drug resistance emergence in *M. tuberculosis* and other pathogens. Prospective cohort studies of participants with and without these polymorphisms should be undertaken to inform clinical trials of personalized pathogen genomic therapy. This novel ancestral state genome-wide survival analysis could also be employed to predict and prevent the emergence of resistance or indeed any important trait of interest in other organisms.

Chapter 4

Within-host diversity improves phylogenetic and transmission reconstruction of SARS-CoV-2 outbreaks

Role in the study: A version of this chapter is publicly available as a preprint in Torres Ortiz et al. (2022). As part of the study, and under guidance of the thesis supervisors, I conceptualised the project, designed the methodology, analysed the clinical and whole-genome sequencing data, and interpreted the results.

Introduction

Understanding who infects whom in an infectious disease outbreak is a key component of infection prevention and control (Didelot et al. 2012). The use of whole-genome sequencing allows for detailed investigation of disease outbreaks, but the limited genetic diversity of many pathogens often hinders our understanding of transmission events (Campbell et al. 2018). As a consequence of the limited diversity, many index case and contact pairs will share identical genotypes, making it difficult to ascertain

who infected whom.

Within-sample genetic diversity is common among a wide variety of pathogens (Mongkolrattanothai et al. 2011; Dinis et al. 2016; Lieberman et al. 2016; Leitner 2019; Popa et al. 2020). As shown in (Fig. 1.3), this diversity may be generated *de novo* during infection, by a single transmission event of a diverse inoculum or by independent transmission events from multiple sources (Worby et al. 2014). The maintenance and dynamic of within-host diversity is then a product of natural selection, genetic drift, and fluctuating population size (Didelot et al. 2012). The transmission of within-host variation between individuals is also favored as a large inoculum exposure is more likely to give rise to infection (Murphy et al. 1984; Han et al. 2019; Sender et al. 2021; Spinelli et al. 2021; Trunfio et al. 2021; Lee et al. 2022). The amount of within-sample diversity transmitted from index-case to contact is determined by the bottleneck size (Zwart and Elena 2015), with stringent bottlenecks limiting the number of genotypes transmitted from the host to the recipient, and wide bottlenecks allowing for the transmission of higher levels of genetic diversity (Worby et al. 2014). Bottleneck size estimation remains challenging, have high uncertainty and often range considerably between studies. For instance, SARS-CoV-2 bottleneck estimates using genetic data, which rely on the number of shared within-host variants between known transmission pairs, range from fewer than 10 virions (Lythgoe et al. 2020; Valesano et al. 2021) to hundreds or even thousands of viral particles (Popa et al. 2020; San et al. 2021). These estimates are likely to differ due to the stochastic nature of transmission, the effect of the different routes of transmission, lack of reliable transmission pairs or epidemiological data, sequencing errors or disparate variant calling pipelines and minor variant thresholds (Braun et al. 2021).

Most genomic and phylogenetic workflows involve either genome assembly or alignment of sequencing reads to a reference genome. In both cases, conventionally the resulting alignment exclusively represents the most common nucleotide at each

position. This is often referred to as the consensus sequence. Although genome assemblers may output contigs (combined overlapping reads) representing low frequency haplotypes, only the majority contig is kept in the final sequence. In a mapping approach, a frequency threshold for the major variant is usually pre-determined, under which a position is considered ambiguous. The lack of genetic variation between temporally proximate samples and the slow mutation rate of many pathogens results in direct transmission events sharing exact sequences between the hosts when using the consensus sequence approach. For instance, the substitution rate of SARS-CoV-2 has been inferred to be around 2 mutations per genome per month (Harvey et al. 2021). Given its infectious period of 6 days (Byrne et al. 2020), most consensus sequences in a small-scale outbreak will show no variation between them. This lack of resolution and poor phylogenetic signal complicate the determination of who infected whom, relying exclusively on epidemiological information.

I hypothesize that the failure of consensus sequence approaches to capture within-sample variation arbitrarily excludes meaningful data and limits the ability to determine who infected whom, and that including within-sample diversity in a phylogenetic model would significantly increase the evolutionary and temporal signal and thereby improve our ability to infer infectious diseases transmission events.

This hypothesis was tested using multi-institutional SARS-CoV-2 outbreaks across London hospitals that were part of the COVID-19 Genomics UK (COG-UK) consortia (COVID-19 Genomics UK (COG-UK) 2020). Technical replicates, repeated longitudinal sampling from the same patient, and epidemiological data allowed us to evaluate the presence and stability of within-sample diversity within the host and in independently determined transmission chains. I also evaluated the use of within-sample diversity in phylogenetic analysis by conducting outbreak and phylogenetic simulations of sequencing data using a phylogenetic model that accounts for the presence and transmission of within-sample variation. I then show the effects on phylogenetic

inference of using consensus sequences in the presence of within-sample diversity, and propose that existing phylogenetic models can leverage the additional diversity given by the within-sample variation and reconstruct the phylogenetic relationship between isolates. Additionally, I performed outbreak simulations. Lastly, I demonstrate that by taking into account within-sample diversity in a phylogenetic model the temporal signal in SARS-CoV-2 outbreak analysis can be improved. Using both phylogenetic outbreak reconstruction and simulation, I show that this approach is superior to the current gold standard whole-genome consensus sequence methods.

Methods

Model for within-host diversity

To test the accuracy of different models at inferring known phylogenies, 100 random phylogenetic trees with 100 tips each were generated using the function *rtree* within the R package *ape* (Paradis et al. 2004). Whole-genome alignments were simulated from the random 100 phylogenies with the function *SimSeq* of the R package *phangorn* (Schliep 2011; R Core Team and R Foundation for Statistical Computing 2021) using a model with 16 character states that represent the combinations of the 4 nucleotides with each other as minor and major alleles (Fig. 4.3d). Three substitution rates for the model were considered: a rate at which minor variants evolve, equal to 1; the rate at which minor variants are lost, leaving only the major nucleotide at that position, equal to 100; and the rate at which minor/major variants are switched, equal to 200.

Two types of alignments were generated from the simulated genomes: a consensus sequence, where only the major allele was considered; and an alignment that retained the major and minor allele information as independent character states. RaxML-NG (Kozlov et al. 2019) was used to infer phylogenetic trees. The consensus sequence

was analyzed with a GTR+ γ model, while the PROTGTR+ γ model was used for the alignment with intra-host diversity and major/minor variant information.

Several metrics were used to compare the 200 inferred phylogenetic trees with their respective starting phylogeny from which the sequences were simulated (Table A4.1). All metrics were available in R and were suitable for unrooted trees, using the option 'rooted=FALSE' where appropriate. The Robinson-Foulds (RF) distance (Robinson and Foulds 1981) calculates the number of splits differing between both phylogenetic trees. For the weighted Robinson-Foulds (wRF), the distance is expressed in terms of the branch lengths of the differing splits. The Kuhner-Felsenstein (KF) distance (Kuhner and Felsenstein 1994) considers the edge length differences in all splits, regardless of whether the topology is shared or not. Last, the Penny-Steel (PS) distance or path difference metric (Steel and Penny 1993) calculates the pairwise differences in the path of each pair of tips, with the weighted Penny-Steel distance (wPS) using branch length to compute the path differences. All functions were used as implemented in the package *phangorn* (Schliep 2011) within R (R Core Team and R Foundation for Statistical Computing 2021).

Outbreak simulations

Disease outbreaks of size between 10 and 15 hosts were simulated using TransPhylo (Didelot et al. 2017), with a mean generation time of 5 days and a mean sampling time of 7 days, both parameters with standard deviation of 1 day (Wang et al. 2020; Hart et al. 2022). To ensure that the outbreak ends, the negative binomial distribution for the offspring number was set with a mean of 1 and a dispersion parameter of 0.5, resulting in a basic reproductive number R_0 of 1. To simplify the simulations, all hosts from the outbreak were sampled. A total of 20 outbreaks were simulated. The population evolution within and between hosts was simulated using fastsimcoal2 (Excoffier et al. 2013) as previously described by De Maio et al. (2018). Sequences

were simulated for a within-host population size of 1000 and a genome size of 1000bp. To understand the effect of transmission bottleneck size in phylogenetic inference, varying values of bottleneck size were used along a constant mutation rate of 5×10^{-6} mutations per base per generation cycle. Additionally, sequences were simulated at different mutation rates with a constant bottleneck size of 10 pathogens. Sequences with the varying bottleneck sizes and mutation rates were simulated using the same 20 simulated outbreaks. Phylogenetic trees were inferred from the alignments using RaxML-NG as previously described. The resulting trees were time-calibrated using the additive uncorrelated relaxed clock model (ARC) as implemented in BactDating (Didelot et al. 2018). The root of the outbreak was inferred as part of the dating model. The inferred trees were compared to the known simulated phylogenies using the Kuhner-Felsenstein distance (Kuhner and Felsenstein 1994).

Amplification and whole-genome sequencing

SARS-CoV-2 real-time qPCR confirmed isolates from London hospitals were collected as part of the routine diagnostic service at Great Ormond Street Hospital NHS Foundation Trust (GOSH) (Storey et al. 2021) and the COVID-19 Genomics UK Consortium (COG-UK) (COVID-19 Genomics UK (COG-UK) 2020) between March and December 2020, in addition to epidemiological and patient metadata (Table 4.1). SARS-CoV-2 whole-genome sequencing was performed by UCL Genomics. cDNA and multiplex PCR reactions were prepared following the ARTIC nCoV-2019 sequencing protocol (Tyson et al. 2020). The ARTIC V3 primer scheme (ARTIC Network 2021) was used for the multiplex PCR, with a 65°C, 5 min annealing/extension temperature. Pools 1 and 2 multiplex PCRs were run for 35 cycles. 5µL of each PCR were combined and 20µL nuclease-free water added. Libraries were prepared on the Agilent Bravo NGS workstation option B using Illumina DNA prep (Cat. 20018705) with unique dual indexes (Cat. 20027213/14/15/16). Equal volumes of the final libraries were pooled, bead purified and sequenced on the Illumina NextSeq 500 platform using

a Mid Output 150 cycle flowcell (Cat. 20024904) (2 x 75bp paired ends) at a final loading concentration of 1.1pM.

Whole-genome sequence analysis of SARS-CoV-2 sequences

Raw illumina reads were quality trimmed using Trimmomatic (Bolger et al. 2014) with a minimum mean quality per base of 20 in a 4-base wide sliding window. The 5 leading and trailing bases of each read were removed, and reads with an average quality lower than 20 were discarded. The resulting reads were aligned against the Wuhan-Hu-1 reference genome (GenBank NC_45512.2, GISAID EPI_ISL_402125) using BWA-mem v0.7.17 with default parameters (Li and Durbin 2010). The alignments were subsequently sorted by position using SAMtools v1.14 (Li et al. 2009). Primer sequences were masked using ivar (Grubaugh et al. 2019).

Single-nucleotide variants were identified using the pileup functionality of SAMtools (Li et al. 2009) via the pysam package in Python (github.com/pysam-developers/pysam). Variants were further filtered using BCFtools (Danecek et al. 2011). Only variants with a minimum depth of 50x and a minimum base quality and mapping quality of 30 were kept. Additionally, variants within low complexity regions identified by sdust (github.com/lh3/sdust) were removed. Problematic sites previously identified by De Maio et al. (2020) were masked to avoid systematic sequencing errors and phylogenetic bias. For positions where only one base was present, the minimum depth was 20 reads, with at least 5 reads in each direction. Positions with low frequency variants were filtered if the total coverage at that position was less than 100x, with at least 20 reads in total and 5 reads in each strand supporting each of the main two alleles.

Two different alignments were prepared from the data. First, an alignment of the consensus sequence where the most prevalent base at each position was kept. Variants

where the most prevalent allele was not supported by more than 60% of the reads were considered ambiguous. Additionally, an alignment reflecting within-sample variation at each position as well as which base is the most prevalent and which one appears at a lower frequency by using the IUPAC nomenclature for amino acids (IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) 1984).

For the two different alignments, maximum likelihood phylogenies were inferred by using RAXML-NG (Kozlov et al. 2019) with 20 starting trees (10 random and 10 parsimony), 100 bootstrap replicates, and a minimum branch length of 10^{-9} . For the consensus sequence, the GTR model was used. For the alignment reflecting within-host diversity, a model with amino acid nomenclature (PROTGTR) was used. All models allowed for a γ distributed rate of variation among sites. Phylogenetic trees were time-calibrated using the known collection dates and the ARC model within BactDating (Didelot et al. 2018).

Results

Sampling, demographics and metadata

Between March 2020 and November 2020, 451 healthcare workers, patients and patient contacts at the participating North London Hospitals were diagnosed at the Camelia Botnar Laboratories with SARS-CoV-2 by PCR as part of a routine staff diagnostic service at Great Ormond Street Hospital NHS Foundation Trust (GOSH). The mean participant age was 40 years old (median 38.5 years old, interquartile range (IQR) 30-50 years old), and 60% of the participants were female (Table 4.1). A total of 289 were whole-genome sequenced using the Illumina NextSeq platform, which resulted in 522 whole-genome sequences including longitudinal and technical replicates. All samples were SARS-CoV-2 positive with real time qPCR cycle threshold (C_t) values ranging from 16 to 35 cycles (Table 4.1). The earliest sample was collected

on 26th March 2020, while the latest one dated to 4th November 2020 (Fig. A4.1a). A total of 291 samples had self-reported symptom onset data, for which the mean time from symptom onset to sample collection date was 5 days (IQR 2-7 days, Fig. A4.1b). More than 90% of the samples were taken from hospital staff, while the rest comprised patients and contacts of either the patients or the staff members (Table 4.1).

Genomic analysis of SARS-CoV-2 sequences

Whole-genome sequences were mapped to the reference genome resulting in a mean coverage depth of 2177x (Fig. A4.2). A total of 454 whole-genomes with mean coverage higher than 10x were kept for further analysis. Allele frequencies were extracted using the pileup functionality within BCFtools (Danecek et al. 2011) with a minimum base and mapping quality of 30, which represents a base call error rate of 0.1%. Variants were filtered further for read position bias and strand bias. Only minor variants with an allele frequency of at least 1% were kept as putative variants. Samples with a frequency of missing bases higher than 10% were excluded, keeping 350 isolates for analysis. The mean number of low frequency variants was 12 (median 3, IQR 1.00 – 9.75), although both the number of variants and its deviation increased at high C_t values (Fig. A4.3).

Within-sample variation is stable between technical replicates

To understand the stability of within-sample variation and minimize spurious variant calls, technical replicates of 17 samples were sequenced and analyzed. Overall, when the variant was present in both duplicates the correlation of the variant frequencies was high ($R^2 = 0.9$, Fig. 4.1a right). The high correlation was also maintained at low variant frequencies (Fig. 4.1a left).

Minor variants were less likely to be detected or shared when one or more of the paired samples had a low viral load. These discrepancies may appear because of

Table 4.1: Sample collection and demographics.

	Number	Percentage
Total samples	451	100
Hospital		
Barnet, Enfield & Haringey Mental Health	18	3.99
Camden and Islington Mental Health	3	0.67
Chase Farm	1	0.22
FLARE Trial	3	0.67
GOSH	182	40.35
North Middlesex University Hospital	105	23.28
Royal Free Hospital	50	11.09
UCLH	10	2.22
Whittington Health	79	17.52
Missing	0	0
Role		
Staff	403	89.36
Patient	14	3.1
Contact	32	7.1
Missing	2	0.44
Sex		
Female	274	60.75
Male	161	35.7
Undertermined	3	0.67
Missing	13	2.88
Age		
(0,10]	8	1.77
(10,20]	9	2
(20,30]	111	24.61
(30,40]	111	24.61
(40,50]	98	21.73
(50,60]	74	16.41
(60,70]	22	4.88
(70,80]	10	2.22
(80,90]	3	0.67
Missing	5	1.11
Ct		
(15,20]	15	3.33
(20,25]	85	18.85
(25,30]	149	33.04
(30,35]	197	43.68
Missing	5	1.11

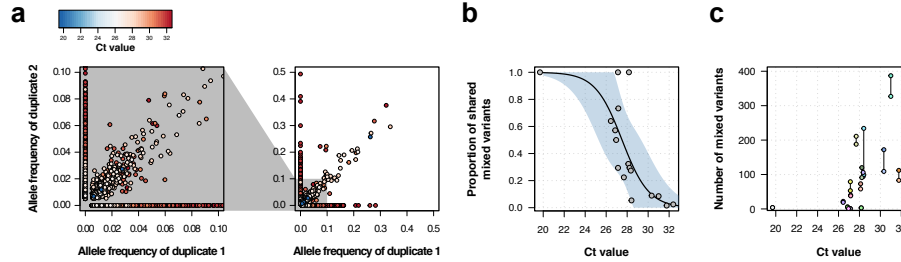


Figure 4.1: Genomic analysis of technical duplicates.

a Allele frequency comparison between technical replicates for all frequencies (right) and for frequencies up to 1% (left). Colors represent the C_t value for the sample. **b** Proportion of shared mixed variants between technical replicates in relation to the C_t value. **c** Total number of mixed variants in relation to the C_t value. Lines linked two technical replicates. Each sequence has a different color, with sequences from the same patient having a different shade of the same color.

amplification bias caused by low genetic material, base calling errors due to low coverage, or low base quality. The mean proportion of discrepant within-sample variants between duplicated samples was 0.39 (sd = 0.29), although this varied between duplicates (Fig. A4.4). C_t values in RT-PCR obtained during viral amplification are inversely correlated with viral load (Tom and Mina 2020). The proportion of shared intra-host variants was negatively correlated with C_t values in a logistic model (estimate=-0.78, p-value=0.008), with higher C_t values associated with a lower amount of shared intra-host variants (Fig. 4.1c). The number of within-sample variants detected also increased with C_t value, as well as the deviation in the number of variants between duplicates (Fig. 4.1d). This could be explained either by an increase in the number of spurious variants at low viral loads (Tonkin-Hill et al. 2021), biased amplification of low level sub-populations minor rare alleles (McCrone and Lauring 2016), or due to the accumulation of within-host variation through time, as viral load (with rising C_t values) also decreases from time since infection.

Based on these results, only samples with a C_t value equal or lower than 30 cycles

were considered, which resulted in 249 samples kept for analysis. For the filtered dataset, 414 out of 29903 positions were polymorphic for the consensus sequence, while the alignment with within-sample diversity had 1039 SNPs. Of these, 699 positions had intra-host diversity, of which 78% (549/699) were singletons. The majority of samples (207/249, 83%) contained at least 1 position with a high quality within-host variant, and the median amount of intra-host variants per sample was 2 (IQR 1-4.5).

Within-sample variation is shared between epidemiologically linked samples

Given the limited genomic information in the consensus sequences, epidemiological data is often necessary to infer the directionality of transmission. The proportion of shared within-sample variants (shared variants divided by total variants in the pair) was calculated by performing a pairwise comparison of all isolates. The comparison was done in samples that a) did not have any recorded epidemiological link, b) samples that were from the same hospital (possibly linked), c) samples that were part of the same department within the same hospital (probable link), and d) samples that had an epidemiological link within the same department of the same hospital (proven link), e) were a longitudinal replicate from the same patient and f) a technical replicate from the same sample.

The concordance between epidemiological and genomic data was tested by determining the genetic distance between pairs of samples with epidemiological links and without them. Pairs of samples from the same hospital, department, epidemiologically linked, or longitudinal and technical replicates were more closely located in the consensus phylogenetic tree than those samples that did not have any relationship, although this difference was small in the case of pairs of samples from the same hospital (Table 4.2).

Table 4.2: Phylogenetic distance (substitutions per genome per year) between pairs of samples.

Sample relationship	Estimate (95%CI)	p-value
None	11.73 (11.63 - 11.83)	Reference
Hospital	10.3 (10.09 - 10.54)	$< 1 \times 10^{-4}$
Department	5.35 (4.21 - 6.49)	$< 1 \times 10^{-4}$
Epidemiological	1.62 (0.53 - 2.72)	$< 1 \times 10^{-4}$
Longitudinal duplicates	0.01 (-1.84 - 1.86)	$< 1 \times 10^{-4}$
Technical replicate	0 (-4.29 - 4.29)	$< 1 \times 10^{-4}$

The proportion of shared within-host variants was significantly higher between technical replicates, longitudinal duplicates, epidemiologically linked samples, and samples taken from individuals from the same department when compared to pairs with no epidemiological links (Fig. 4.2). The probability of sharing a low frequency variant was inferred using a logistic regression model (Fig. A4.5). There was a tendency for the probability to increase with variant frequency, but the association was not strong (Odds ratio 1.8, 95% CI 0.9 – 3.5, $p=0.08$). The probability of sharing a variant for samples with no epidemiological links was 9.5×10^{-6} (95% CI 8.8×10^{-6} – 1.02×10^{-5}). Samples from the same hospital did not have a probability significantly higher than those without any link (3.3×10^{-3} , 95% CI 2.7×10^{-3} – 4.03×10^{-3}). On the other hand, pairs from the same department, with epidemiological links, replicates or technical replicates all had a higher probability of sharing a low frequency variant when compared to those pairs with no link (all Wald test p-values < 0.001). The inferred probabilities for pairs from the sample department was 1.4% (95% CI 0.9% – 2.1%), which increased to 5% for pairs with epidemiological links (95% CI 4.2% – 6.4%). For longitudinal replicates, the probability was inferred to be 38% (95% CI

35% – 41%), while technical replicates were estimated to have the highest probability (70%, 95% CI 64% – 76%).

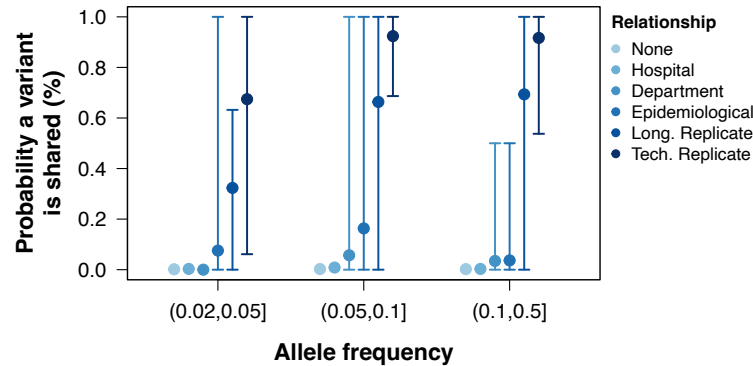


Figure 4.2: Probability of sharing within-host variants in sample pairs.

The probability of variants shared between pairs of samples calculated as the number of low frequency variants in both samples divided by the total number of variants between the pair. Colors grouped samples by their relationship. Points represent the mean probability a variant is shared between all pairwise samples within a group and allele frequency. Error bars show the 95% CI.

Within-host diversity model outperforms the consensus model in simulations

The effect of within-sample diversity in phylogenetic inference was tested by evaluating the accuracy in the reconstruction of known phylogenetic trees using a conventional phylogenetic model and a model that accounts for within-sample variation.

The presence of within-sample diversity was coded in the genome alignment using existing IUPAC nomenclature (IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) 1984). For the consensus sequence alignment, only the 4 canonical nucleotides were used (Fig. 4.3a,b), while the proposed alignment retained the major

and minor allele information as independent character states (Fig. 4.3c,d).

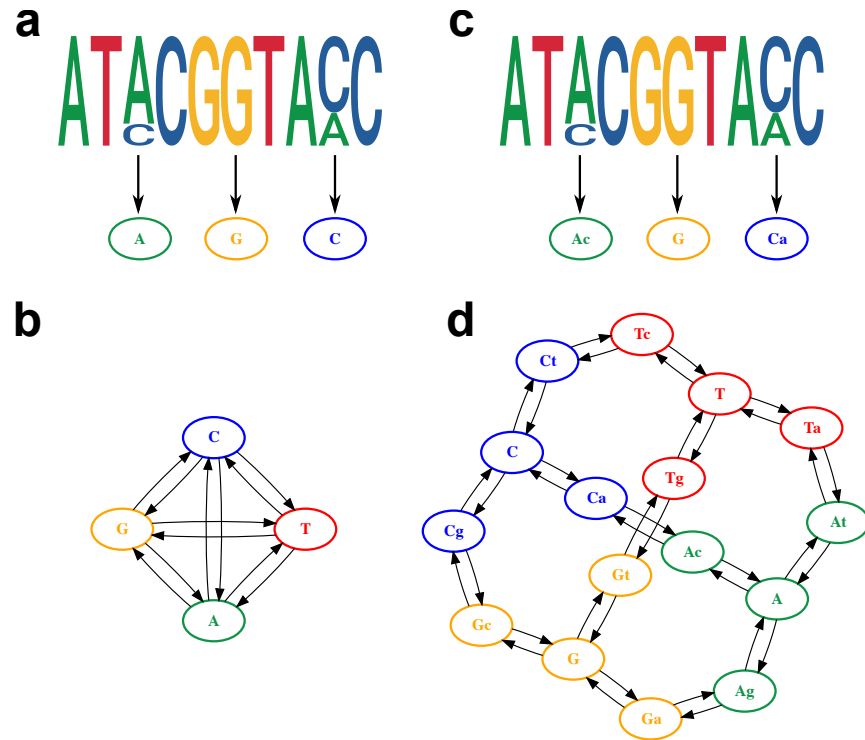


Figure 4.3: Model of within-host diversity.

Proposed evolutionary model of within-host diversity in genomic sequences. Uppercase letters represent the major variant in the population, while lowercase letters indicate presence of a minor variant alongside the major one. **a, c** Genome sequences where some positions show within-sample variation (top), represented by a major allele (big size letter) and a minor one (smaller size), as well as its representation in the alignment (bottom). **b, d** Models of nucleotide evolution. Character transitions are indicated by arrows. **a** Consensus sequence, where only the major allele is represented in the alignment. **b** Model of nucleotide evolution using the consensus sequence, with four character states representing the four nucleotides. **c** Sequence with within-sample variation, represented by an uppercase letter for the major allele and a lower case letter for the minor allele. **d** Model of nucleotide evolution with 16 character states accounting for within-sample variation.

In order to evaluate the bias in tree inference with and without the inclusion of within-sample diversity, genome alignments for 100 random trees were simulated using a phylogenetic model where both major and minor variant combinations were considered, resulting in a total of 16 possible states (Fig. 4.3d) and the substitution rates shown in Table A4.2. From the simulated genomes, two types of alignments were generated: a consensus sequence, where only the major allele was considered (Fig. 4.3a); and an alignment that retained the major and minor allele information as independent character states (Fig. 4.3c). From the simulated alignments, RaxML-NG was used to infer phylogenetic trees (Kozlov et al. 2019). The consensus sequence was analyzed with a GTR+ γ model, while the PROTGTR+ γ model was used in order to accommodate the extra characters of the alignment with within-sample diversity and major/minor variant information.

The two models were evaluated for their ability to infer the known phylogeny that included within-host diversity. The estimated phylogenies were compared to the known tree using different measures to capture dissimilarities in a variety of aspects relevant to tree inference (Table A4.1). For all the metrics employed, the phylogenies inferred explicitly using within-host diversity as independent characters approximated better to the initial tree than the one using the consensus sequence (Fig. 4.4). Additionally, the transition/transversion rates inferred by the phylogenetic models accounting for within-host diversity accurately reflect the rates used for the simulation of genomic sequences (Table A4.2, A4.3 and A4.4).

The effect of within-host diversity in phylogenetic inference was further assessed by simulating pathogen evolution throughout the time frame of infectious disease outbreaks. Outbreak simulations were performed using TransPhylo (Didelot et al. 2017) with a host population varying between 10 and 15 hosts, no recombination, complete sampling of the outbreak and selecting epidemiological parameters to match the transmission dynamics of SARS-CoV-2. For each outbreak, the evolution and

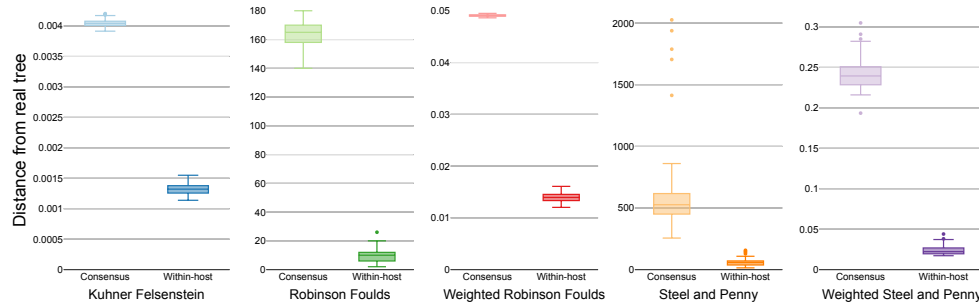


Figure 4.4: Similarity scores for inferred trees.

Comparison of the phylogenetic trees inferred using three simulated sequences and different phylogenetic models with the known starting tree. Colors differentiate the metrics used for the comparison.

transmission of the pathogen population within each host was simulated with varying values of mutation rates and transmission bottlenecks using fastsimcoal2 (Excoffier et al. 2013) as previously described by De Maio et al. (2018). The resulting phylogenetic trees were compared to the real outbreak phylogeny using the Kuhner-Felsenstein (KF) distance (Kuhner and Felsenstein 1994). Even though using consensus sequences performed better than a random distribution of trees, using within-host diversity outperformed the consensus sequence in all instances (Fig. 4.5). The phylogenies inferred using within-host diversity were more similar to the real outbreak phylogeny for wider bottleneck sizes, with the best performance when no bottleneck was present. As expected, both the consensus sequences and the sequences reflecting within-sample diversity were more informative at higher mutation rates, even though the consensus sequence only showed improvement with a mutation rate of 10^{-3} mutations per base per generation cycle (Fig. 4.5).

Within-host diversity improves the resolution in SARS-CoV-2 phylogenetics

Genome sequences collected at different time points are expected to diverge as time progresses, resulting in a measurable positive correlation between the isolation date

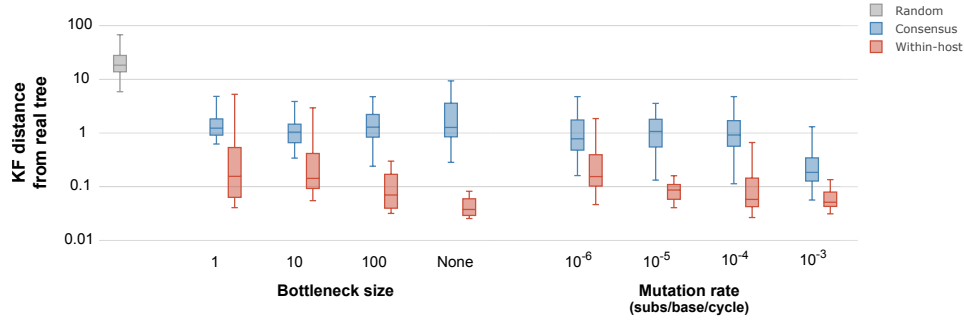


Figure 4.5: Inferred phylogenetic trees from outbreak simulations.

Kuhner and Felsenstein (KF) tree distance between phylogenetic trees from simulated outbreaks. Phylogenies were inferred using consensus sequences (blue) and alignments reflecting within-sample diversity (red), and compared to the known phylogeny of the simulated outbreak. For reference, grey color represents a set of random trees. Outbreak simulations were performed with different bottleneck sizes and mutation rates. The mutation rate is measured as the number of mutations per base per generation cycle.

and the number of accumulated mutations, also known as temporal signal (Rieux and Balloux 2016). The alignment with consensus sequences and the one reflecting within-sample variation were used to infer two different phylogenetic trees (Fig. A4.6). Longitudinal samples in the phylogeny inferred using within-host diversity reflected the expected temporal signal, with an increase in genetic distance as time progressed between the longitudinal pairs in a linear model (coefficient 2.24, 0.59 - 3.88 95% CI, $p = 0.019$, Fig. A4.7). The difference in C_t value among longitudinal duplicates was not correlated with higher genetic distances (coefficient 1.62, -0.66 - 3.91 95% CI, $p = 0.2$).

The impact of using within-sample variation on the temporal structure of the phylogeny was analyzed by systematically identifying clusters of tips in the phylogenetic tree with an identical consensus sequence and no temporal signal. A root-to-tip analysis was then performed using the tree inferred with intra-sample diversity. Only clusters with more than 3 tips were used for the root-to-tip analysis. The majority of clusters (10/11) showed a positive correlation between the distance of the tips to the

root and the collection dates, demonstrating a significant temporal signal between samples when there was none using the conventional consensus tree (Fig. 4.6).

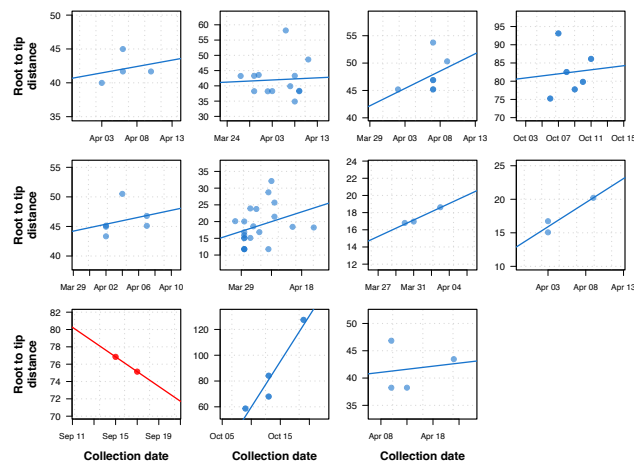


Figure 4.6: Previously uninformative clusters present temporal signal when using within-sample diversity.

A set of 11 outbreak clusters (one per panel, each plotting the root to tip distance against time) in which all samples had identical consensus genomes sequences (and therefore no temporal signal). Blue colors indicate those regressions that after utilizing within sample diversity now have a positive slope (temporal signal), and red shows those regressions that have a negative slope (misleading or false positive temporal signal).

To illustrate the downstream application of the improved phylogenetic resolution, a time-calibrated phylogeny with the collection dates of the tips was inferred using BactDating (Didelot et al. 2018) (Fig. A4.8) and calculated the likelihood of transmission events within potential epidemiologically identified outbreaks using a Susceptible-Exposed-Infectious-Removed (SEIR) model (Eldholm et al. 2016). The SEIR model was parameterized with an average latency period of 5.5 days (Xin et al. 2022), an infectious period of 6 days (Byrne et al. 2020), and a within-host coalescent rate of 5 days as previously estimated for SARS-CoV-2 (Wang et al. 2020). The likelihood of transmission was calculated for every pair of samples, while the Edmonds algorithm as implemented in the R package *RBGL* (Carey et al. 2021) was used to infer the graph with the optimum branching (Fig. 4.7c,d; Fig. A4.9).

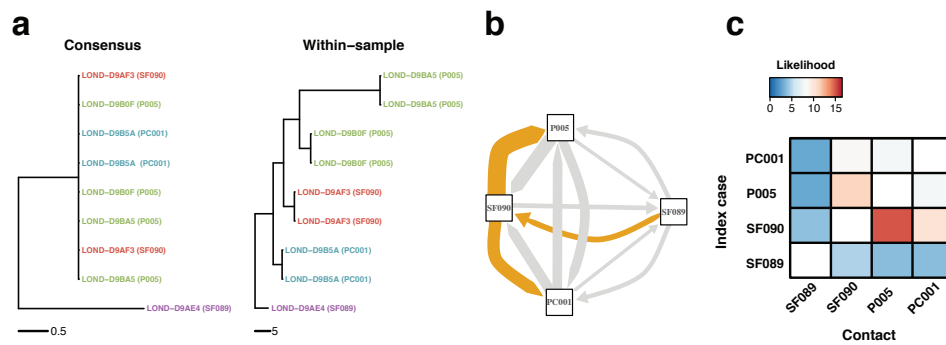


Figure 4.7: Within-sample variation improves resolution of infectious disease outbreaks.

Effect of using low frequency variants in phylogenetic inference. **a** Maximum likelihood phylogeny using the consensus sequences (left) and the alignment leveraging within-sample variation. Replicates of the same sample share the same color. Sample IDs are coded as follows: SF, for staff members; P, for patients; and PC, for patient contacts. **b** Transmission network inferred using within sample variation. Edge width is proportional to the likelihood of direct transmission using a Susceptible-Exposed-Infectious-Removed (SEIR) model. Colored edges represent the Edmunds optimum branching and thus the most likely chain. **c** Heatmap of the likelihood of direct transmission between all pairwise pairs of samples using a SEIR model. Vertical axis is the infector while the horizontal axis shows the infectee.

Discussion

Detailed investigation of transmission events in an infectious disease outbreak is a prerequisite for effective prevention and control. Although whole-genome sequencing has transformed the field of pathogen genomics, insufficient pathogen genetic diversity between cases in an outbreak limits the ability to infer who infected whom. Using multi-hospital SARS-CoV-2 outbreaks and phylogenetic simulations, I show that including the genetic diversity of subpopulations within a clinical sample improves phylogenetic reconstruction of SARS-CoV-2 outbreaks and determines the direction of transmission when using a consensus sequence approach fails to do so.

The majority of samples sequenced harbored variants at low frequency that remained stable in technical replicates. However, within-host variation was less consistent between paired samples with a lower viral load (higher C_t). This is likely to be a consequence of low starting genetic material giving rise to amplification bias during library preparation and sequencing. Establishing a cut-off for high C_t values is therefore important to accurately characterize within-host variation. For this chapter, samples with a C_t value higher than 30 cycles based on the diagnostic PCR used at GOSH were excluded. Since C_t values are only a surrogate for viral load and are not standardized across different assays (Evans et al. 2021), appropriate thresholds would need to be determined for other primary PCR testing assays.

The generation, maintenance and evolution of subpopulations within the host reflect evolutionary processes which are meaningful from phylogenetic and epidemiological perspectives. Subpopulations within a host can emerge from three mechanisms: de novo diversification in the host, transmission of a diverse inoculum, or multiple transmission events from different sources. If the subpopulations are the result of de novo mutations, nucleotide polymorphisms within the subpopulations accumulate over time and may therefore result in a phylogenetic signal useful for phylogenetic

inference. In the data used for this chapter, longitudinal samples taken at later time points were demonstrated to accrue genomic variation. Although this pattern can be confounded by decreasing viral load as infection progresses, C_t values in this dataset were not correlated with a higher genetic distance, and clusters in the data containing both longitudinal and technical replicates also corroborate these results. Transmission of a diverse inoculum also gives rise to phylogenetically informative shared low frequency variants, as the results in this chapter show that transmission pairs are more likely to share variants at low frequency. The effect of multiple transmission events in the phylogeny depends on the relatedness of both index cases and the bottleneck size in each transmission event.

Paired samples with epidemiological links and from the same department shared a higher proportion of low frequency variants and were located closer in the consensus tree than samples with no relationship. Similarly, samples with shorter distance in the consensus phylogeny were more likely to share low frequency variants. These patterns suggest that the distribution of low frequency variants is linked to events of evolutionary and epidemiological interest. The fact that technical duplicates shared more within-host diversity than longitudinal replicates of the same sample suggests that much of the variation within hosts is transitory. Therefore, within-host diversity may be relevant on relatively short time scales, which is precisely where consensus sequences lack resolution. Combining the data derived from fixed alleles in the consensus sequences and transient within-sample minor variation enables an improved understanding of the relatedness of pathogen populations between hosts.

The effects of neglecting within-host diversity in phylogenetic inference were analyzed by using simulated sequences under a phylogenetic model that reflects the presence and evolution of within-host diversity. I compared a conventional consensus phylogenetic model and a model that leverages within-sample diversity, and evaluated their ability to infer the known phylogeny. The phylogenetic model proposed in this

chapter incorporates within-sample variation by explicitly coding major and minor nucleotides as independent characters in the alignment. In this chapter, it is shown that phylogenies inferred using the conventional consensus sequence approach were heavily biased and unrepresentative of the known structure of the simulated tree. However, sequences that included within-host diversity were shown to infer less biased phylogenetic trees.

The phylogenetic simulations were complemented with tree inference of outbreaks simulated using TransPhylo (Didelot et al. 2017). The simulations were parameterized to reflect the transmission dynamics of SARS-CoV-2, including a generation time of 5 days and a sampling time of 7 days. Given this parameters, most simulated outbreaks lasted less than a month. Since TransPhylo simulates phylogenetic trees alongside the outbreak simulation, I could directly compare the inferred phylogenies with the known simulated trees. Genetic sequences within the outbreak were then simulated using fastsimcoal2 (Excoffier et al. 2013) as previously described by De Maio et al. (2018). Using a mutation rate of 5×10^{-6} mutations per base per replication cycle, as previously calculated for SARS-CoV-2 and other betacoronaviruses (Sender et al. 2021; Amicone et al. 2022), and varying bottleneck sizes, I showed that tree inference using within-sample diversity improves as the transmission bottleneck widens, although even at low bottleneck sizes trees inferred using within-sample diversity are more accurate than those inferred using consensus sequences. Similarly, using varying mutation rates and a constant bottleneck size of 10 pathogens, I showed that tree inference was more accurate as mutation rates increased, although inference using consensus sequences improved only at a very high mutation rate of 10^{-3} mutations per base per cycle, which has only been observed in some HIV studies (Cuevas et al. 2015) and RNA viruses (Fig. 1.1). Together, the simulations show that at the short time frame of disease transmission, phylogenetic inference using alignments that contain information regarding within-sample diversity outperform phylogenies inferred with consensus sequences, even at narrow transmission bottlenecks and very

low mutation rates.

Previous studies have addressed the use of within-host variation to infer transmission events. Wymant et al. (2018b) employed a framework based on phylogenetic inference and ancestral state reconstruction of each set of populations detected within read alignments using genomic windows. This chapter extends the previous work by coding genome-wide diversity within the host directly in the alignment and the phylogenetic model. (De Maio et al. 2018) proposed direct inference of transmission from sequencing data alongside host exposure time and sampling date within the bayesian framework BEAST2 (Bouckaert et al. 2014). The approach proposed in this chapter is focused on directly improving the temporal and phylogenetic signal of whole-genome sequences, and it's especially suited for use in applications and analysis that employ a phylogenetic tree as input to infer transmission (Didelot et al. 2017).

Future work will extend this model by including allele frequency data in addition to independent characters for major and minor variants. Phylogenetic models that explicitly include dynamics of within-sample variation and sequencing error may further improve phylogenetic inference or allow researchers to better estimate parameters of interest, including R_0 , bottleneck size, transmissibility and the origin of outbreaks.

This study benefited from the availability of sequenced technical replicates that enabled us to distinguish genuine variation from sequencing noise, especially at low variant frequencies. Similarly, access to longitudinal samples from the same patient allowed us to characterize the spectrum of within host variation and therefore reconstruct transmission chains with more precision.

In line with conventional consensus sequencing approaches, a reference sequence was used for genome alignment and variant calling. Although widely used, one limitation of this approach is a potential mapping bias causing some reads to reflect

the reference base at low frequencies at a position where only a variant should be present. Although stringent quality filtering was applied, the persistence of some false positive minor variants cannot be ruled out. Using genome graphs to map to a reference that encompasses a wider spectrum of variation may alleviate this problem, and could be an interesting addition to pathogen population genomic analysis.

A limitation of the proposed approach is that it can only accommodate two low frequency variants at each position. Even though considering more than two alleles could add more genetic information for phylogenetic inference, the low frequency in the population of most alleles results in most positions being represented by at most two variants, and therefore in practice would have little difference in the analysis. Moreover, the low frequency of a third allele could result in more sequencing and mapping errors which could in turn bias phylogenetic inference and genomic analysis.

The results presented in this chapter demonstrate that within-sample variation can be leveraged to increase the resolution of phylogenetic trees and improve our understanding of who infected whom. Using SARS-CoV-2 hospital outbreaks and simulations, I show that variants at low frequencies are stable, phylogenetically informative and are more often shared among epidemiologically related contacts. By coding within-sample variation directly in the alignment, the additional genetic information can be easily incorporated in phylogenetic inference, facilitating its application within existing epidemiology pipelines and public health infrastructure. I propose that pathogen phylogenetic models should accommodate within-host variation to improve the understanding of infectious disease transmission and aid infection control measures.

Chapter 5

Discussion

As whole-genome sequencing technologies are becoming increasingly cost-effective, have higher-throughput, and are more rapid and accurate, their potential to transform routine clinical microbiology has become more promising. In this thesis, I introduced methodological and conceptual advancements that have a direct impact in the application of genomics in clinical microbiology, especially in phenotype testing, antimicrobial resistance, epidemiology and within-host diversity. The results presented in this thesis are focused on *M. tuberculosis* and SARS-CoV-2, but all the methods and methodological improvements are highly applicable to other infectious diseases.

Genomic analysis of pathogen sequences

Advances in sequencing technologies have made high-throughput genome sequencing available to many clinical microbiology laboratories. However, the standardization of data analysis and bioinformatic workflows is still a challenge, which hinders the full and seamless implementation of genomic pipelines within clinical microbiology. In chapter 2, I first explored biases that can arise when choosing a reference genome, and proposed a bioinformatic pipeline that can ameliorate them.

Using a common reference genome is necessary not only for comparative studies in basic research, but also for many applications within routine clinical microbiology. Testing clinical samples for phenotypes of clinical interest are likely to require alignment of sequencing reads to a well annotated and complete reference assembly. Moreover, aligning genetic sequences within the same genomic coordinates is a prerequisite for genomic epidemiology, particularly for applications based on phylogenetic approaches. As discussed in chapter 2, most bioinformatic pipelines rely on direct mapping of short-reads to a complete reference assembly, introducing inaccuracies in hard to map genomic regions —such as repetitive and low complexity sequences— and when the clinical isolates are distinct from the target reference. *M. tuberculosis* represents a good working example of reference bias. First, the *M. tuberculosis* genome contains many areas that are challenging to resolve by standard methods, especially the *pe/ppe* genes and many paralogue genes such as lipoproteins that show high sequence similarity and are therefore also prone to mapping biases. Moreover, *M. tuberculosis* also presents multiple lineages and sublineages which can introduce mapping artifacts. *M. tuberculosis* has a well annotated and widely used reference genome, H37Rv, a lineage 4 assembly that is ubiquitous in most research projects, irrespective of the lineage of the clinical isolates under study. In chapter 2, I suggest a bioinformatic workflow that incorporates a step of *de novo* assembly to tailor a sample-specific reference genome, against which short-reads are mapped. When employing the suggested workflow, lineage and gene dependent mapping bias were reduced. This improved methodology can be applied to any pathogen, although those species with complex genomes, hypervariable sites or with high genetic diversity between subspecies would be the ones that would benefit the most from this workflow. For instance, due to the low genetic diversity in SARS-CoV-2, the existence of a genetically similar reference genome and the lack of long genomic regions with high mapping errors, I chose to not apply this approach in chapter 4.

There is increasing evidence of the potential role in virulence of the *pe/ppe* gene

family (Fishbein et al. 2015; Ates et al. 2018) and other poorly resolved regions of the *M. tuberculosis* genome such as lipoproteins or ESX secretion systems (Forrellad et al. 2013). Future markers of pathogenicity and other phenotypes of clinical interest in *M. tuberculosis* and other pathogens will inevitably require that genomic regions difficult to map are first fully resolved. The bioinformatic workflow proposed in chapter 2 allows isolates to be analyzed within the same genomic coordinates of the reference assembly, while reducing bias caused by species diversity and hypervariable regions, thus including a higher number of polymorphisms for comparative studies such as GWAS. Moreover, transmission and epidemiological studies that rely on genetic data will also benefit from the variation added. As longer read technologies become more common, repetitive and highly divergent regions will be better resolved. However, the benefits of using a reference tailored to the clinical isolate shown in chapter 2 are likely to remain even as sequencing technologies improve.

Genetic markers of pre-resistance

In the second part of this thesis, the concept of pre-resistance was introduced to describe the effect that the genetic background of drug-susceptible pathogens has on the risk of future drug-resistance acquisition. To do that, chapter 3 describes novel methodological applications based on survival analysis of phylogenetic trees, and introduces theoretical advancements that will improve our understanding of drug resistance acquisition and the relationship between pathogen genotype and the risk of drug resistance acquisition, allowing prediction of drug resistance emergence from genetic data.

In chapter 3, I aimed to detect associations between pathogen phenotypes and genetic polymorphisms within a time-to-event phylogenetic framework, where time is represented by the branch length between internal nodes of the phylogeny and the event refers to the first acquisition of the phenotype along internal branches. The advantage

of the time-to-event method over traditional associations analysis based on linear or logistic models is that it allows to include temporal data, thus potentially increasing the power to detect genotype-phenotype associations when compared to models based on binary data. Moreover, analyzing the ancestral evolution of the phenotype of interest along the phylogenetic tree offers an straight-forward method to account for cryptic population structure, which is a growing concern in association analysis for many pathogens. However, the proposed methodology also has disadvantages. As many phylogenetic based methods, the results may be affected by differential sampling among clades which can result in poor inference of ancestral states and biased branch length calculation. Additionally, although using the first acquisition of the phenotype allows to correct for population structure, it also rapidly reduces the sampling size.

Using the proposed method, I estimate an association between pathogen genotype and the risk of drug resistance acquisition. First, lineage 2 of *M. tuberculosis* showed a higher risk of acquiring resistance than lineage 4 in our dataset of Lima (Peru) and in a publicly available dataset in Samara (Russia). Additionally, a higher risk of acquiring rifampicin resistance was estimated for those strains that are already resistant to isoniazid. Finally, I performed a GWAS to identify genes and polymorphisms in drug susceptible strains associated to a higher risk of drug resistance acquisition.

The methodology described in chapter 3 can be translated to other pathogens in order to establish pre-emptive markers of drug resistance. Recently, Jangir et al. (2022) showed that pre-existing genetic polymorphisms influence the evolution of resistance to antibiotics in *E. coli*. Moreover, this method could also be applied to predict the emergence of other traits of clinical interest, such as virulence or transmission potential.

To understand the effects of pre-resistance mutations, we studied the dynamics of

resistance acquisition using consensus sequences, which reflect exclusively the most prevalent population in the isolated sample. As described in chapter 1 and shown in chapter 4, some pathogens are characterized by large amounts of within-host diversity. Moreover, diversifying selection within the host is favored for some phenotypes. Pre-resistance genotypes could be present in more patients than the ones reported in the chapter 3 dataset if they remain polymorphic at low frequencies within hosts. Deep sequencing at high depth is required to understand the within-host evolution and transmission of pre-resistance genotypes.

The results from chapter 3 not only broaden our understanding of the evolution of drug resistance *M. tuberculosis*, but they can also be applied to develop genetic markers to pre-empt the emergence of drug resistance and implement measures before it occurs. As our repertoire of genetic markers grows and becomes more sophisticated to include phenotypes ranging from drug resistance to transmission and virulence, genetic markers of pre-resistance can also be used to inform treatment options. Understanding which susceptible pathogens have greater risk of acquiring drug resistance could allow drug resistance to be prevented by employing broader treatment regimens to those pathogens at greatest risk of developing resistance.

Further association studies that incorporate a higher number of isolates and more genetic diversity need to be undertaken to confirm the findings of this chapter and to uncover more genetic markers of pre-resistance. Laboratory confirmation, such as in fluctuation assays, also needs to be designed in order to corroborate the polymorphisms found in the phylogenetic analysis. Moreover, the phylogenetic analysis used for the analysis can reveal trends of drug resistance acquisition throughout evolutionary time, which may not correspond to the dynamics of drug resistance during the time frame of disease. Therefore, prospective cohort studies with and without these polymorphisms are required to determine the effects of these mutations during infection, and to inform clinical trials of personalized pathogen genomic therapy.

Implications of within-host pathogen diversity

In the final chapter of this thesis, I analyzed the consequences that within-host diversity have in pathogen phylogenetic inference and show that within-host heterogeneity can be incorporated into phylogenetic analysis to improve tree and transmission inference.

Using phylogenetic and transmission simulations, chapter 4 describes the effects on tree inference when pathogen within-host diversity is present. Moreover, using within-host variation in phylogenetic inference of SARS-CoV-2 hospital outbreaks showed the correlation between the additional diversity added by the within-sample heterogeneity and a better temporal structure in the tree.

SARS-CoV-2 represents a good example to analyze the effects of within-host diversity in phylogenetic inference, given that its slow mutation rate precludes transmission inference using genetic data due to the low genomic diversity accumulated during the time span of disease outbreaks. Furthermore, populations resulting from direct transmission are not expected to diverge considerably in SARS-CoV-2 given the short time between infection and sampling. However, the short incubation and infectious time of SARS-CoV-2 also results in some index cases not accumulating enough measurable within-host diversity. Moreover, the amount of diversity transmitted from index case to recipient depends on the mode of transmission and the type of contact. Therefore, the benefits of using within-host diversity depend on many factors, and some transmission cases cannot be fully resolved even when using within-sample diversity.

The use of within-host diversity for phylogenetic and transmission inference is applicable to other pathogens, although as described above its efficacy will depend on multiple factors. Loose bottlenecks promote the transmission of more low frequency

variants, and therefore pathogens that have routes of transmission that favor wide bottlenecks, such as transmission by direct contact as opposed to aerosol (Varble et al. 2014; Frise et al. 2016), would benefit the most from including within-host diversity in transmission inference. However, the simulations presented in chapter 4 show that even with a bottleneck size of one virion, alignments using low frequency variants outperforms consensus sequences in transmission inference. This is explained by the fact that even when one viral particle is transmitted, using low frequency variants increases the probability that the variant transmitted is reflected in the alignment.

Given the low substitution rate of *M. tuberculosis*, many isolates share the same consensus sequences, which precludes phylogenetic and transmission inference without the use of epidemiological data. Applying the methodology described in chapter 4 could resolve many outbreaks and transmission events. *M. tuberculosis* shows a highly variable but relatively long incubation period that ranges from several weeks to up to a few years (Behr et al. 2018). This long incubation period could favor the establishment of high levels of within-host diversity in the index case, which would support the use of low frequency variants to infer transmission. However, the long incubation time in the transmission recipient could similarly provide enough time for the subpopulations to differ significantly at the time of sampling from those at the time of transmission, decreasing the similarity between index case and recipient. Longitudinal studies in *M. tuberculosis* are required to understand the within-host evolution during infection, while household studies could determine the degree of within-host diversity that is transmitted between known contacts. Moreover, the methodology employed to isolate the bacilli will also affect the amount of within-host diversity available for sequencing. Culture methods typically employed in *M. tuberculosis* decrease the genetic diversity of the sample when compared to sequencing directly from sputum (Nimmo et al. 2019). Similarly, *M. tuberculosis* populations isolated from one sample of sputum may not be fully representative of the diversity within the lung. Studies with multiple biological and technical replicates are thus

required to measure the reproducibility of the genetic diversity obtained with different sampling and extraction methods.

Even though in chapter 4 I showed the benefits of using within-host diversity in phylogenetic and transmission inference, its inclusion in routine public health and microbiology remains a challenge. Measurements of within-host diversity are very sensitive to variant calling pipelines and errors (McCrone and Luring 2016), as well as sequencing technology (Grubaugh et al. 2019), which hampers the standardization and use of within-host diversity outside of research studies. The amount of within-host diversity available to study also depends on sampling strategies and sequencing methodology, with culture characterized by less within-sample diversity than sequencing directly from clinical isolates. Moreover, standardization is challenging as sequencing bias in within-host diversity is inversely correlated to viral load, which in turn varies throughout infection. These challenges may have minimal impact since within-host diversity is most informative in short time scales and within outbreaks, which results in most samples being processed and sequenced within the same laboratory and standards.

In order to incorporate within-host diversity into phylogenetic inference, positions in the genome where populations within the host diverged were indicated using the IUPAC ambiguity codes to represent the majority and minority base, resulting in 16 possible character combinations. This format was chosen as it allows for a seamless implementation into phylogenetic and genomic pipelines and it is efficient for storing large alignments. However, using an alignment limits the analysis to only two possible bases at each position and it cannot incorporate base frequency information. An advantage of incorporating within-host diversity information directly in the alignment is that many genomic repositories for pathogens, such as GISAID (gisaid.org), exclusively share assemblies. Currently, the information regarding sub-populations within each sample is lost, and coding the within-host diversity in the

assembly could incorporate that information without changing existing databases.

The phylogenetic method used in chapter 4 incorporates the within-sample heterogeneity by coding the subpopulations at each position as individual characters. Further phylogenetic methods could incorporate within-host evolution in the tree inference, as well as bottleneck estimates and sequencing error distributions.

Future perspectives

The work presented in this thesis improves our knowledge of the evolution of pathogen phenotypes of clinical interest, and presents methodological advances to apply whole-genome sequencing into clinical microbiology and public health. The publication of this thesis coincides with a higher implementation of genome sequencing data into routine clinical microbiology practice. In this section, I will outline future directions of the work presented in this thesis as well as of clinical microbiology as a field.

Genome-wide associations studies (GWAS) have resulted in a catalogue of genetic polymorphisms in many pathogens associated with antimicrobial resistance, which has fueled the use of molecular markers for drug susceptibility testing. However, there are still mechanisms of resistance not predicted by current genomic signatures, which prompted Ellington et al. (2017) to conclude in a recent review that there is currently insufficient evidence to support the use of whole-genome sequencing to perform drug susceptibility testing to guide clinical decisions, albeit the authors highlighted the near perfect correlation between *rpoB* mutations and rifampicin resistance in *M. tuberculosis*, and emphasized that the application of whole-genome sequencing to infer drug susceptibility should be a funding priority. Further research will be necessary to include non-specific mechanisms of resistance, such as membrane

permeability and regulation of efflux pumps, and antimicrobial resistance caused by changes in transcriptomics or epigenomics. Prediction of other phenotypes such as virulence or transmissibility is still not as advanced as that of antimicrobial resistance. In this thesis, I also argue that genetic markers of pre-resistance that allow researchers to predict the emergence of antimicrobial resistance before it occurs could also be included in the future as they will allow to influence patient treatment and establish a less reactive course of therapy.

The discovery of new phenotypes of clinical interest and the consolidation of the mechanism of existing ones may require a change in genomic research practices. Genomics is currently embarked in a data generation paradigm, where millions of pathogen genomes are sequenced every year with some form of accompanying meta-data to perform association and observational studies. This strategy has led to a multitude of successful hypothesis-free and purely observational projects, including GWAS that have aid defining consistent markers of drug resistance. However, there is an increasing need to couple genomic data with well controlled experiments and a strong theoretical backbone in order to confirm many of the associations found in hypothesis-free studies and described the underlying causal mechanisms of observational patterns (Smith and Cordes 2019).

Additionally, as data sets grow in size, it is important to set sampling and processing standards to avoid falling into some of the big data paradoxes: spurious and coincidental associations are more likely to be found in large data sets (Smith 2020), and at the same time the overconfidence of biased results increases as biased data sets grow in size (Meng 2018).

The pattern mentioned above is exacerbated by the fact that genomic associated metadata is often not publicly available or it is challenging to collect and interpret (Black et al. 2020; Azarian 2022). The increasing use of pathogen whole-genome

sequencing data in reference laboratories and public health agencies has resulted in large amounts of genomic data that can be used for epidemiological and public health studies. However, even though genetic sequences are systematically deposited in public repositories (Balloux et al. 2018), the usefulness of this data relies on the availability of the genomic data alongside with associated metadata—including isolation date, geographical location, as well as pathogen phenotypes and host characteristics and clinical data—which is often not released (Azarian 2022). Moreover, genomic data is often shared in formats that preclude its use for some analysis. For instance, sequences deposited in GISAID during the COVID-19 pandemic were invaluable to understand broad geographical patterns of SARS-CoV-2 transmission, but the database is exclusively based on genome assemblies which prevented the quality analysis of the data or its use in analysis that required other formats such as the study of within-host diversity. Therefore, the future of the field relies on standardized and transparent data sharing protocols that also ensure ethical practices that protect patient clinical data and rewards data generating institutions accordingly.

Another challenge in pathogen genomics is the characterization of within-host diversity. In this thesis, I have explored how the sequencing and use of multiple populations within the host can improve phylogenetic and transmission inference, but future projects will require a better characterization of the population within the host as well as controlled environments and interdisciplinary approaches that merge evolutionary biology, ecology and population genetics. Controlled transmission studies in animal models coupled to deep whole-genome sequencing will allow us to understand the dynamics of within-host diversity during transmission, accurate bottleneck size estimates in different conditions and the capability of whole-genome sequencing to capture transmission dynamics. Moreover, deep sequencing in observational clinical studies with extensive metadata could shed light on the clinical consequences of within-host diversity. The study of heteroresistance is also a key element for future infection control. Many reinfection cases of drug resistant infections can be explained

by the relapse of an already resistant subpopulation, and thus the effect of within-host diversity in drug resistance emergence is essential for the design of control measures and patient care. Given that heteroresistance is often characterized by a resistant population coexisting within the host at a frequency well below 1%, new sequencing protocols will need to be developed to detect variants at that frequency and, possibly, directly from clinical samples.

Finally, the effective inclusion of pathogen genomics within public health requires that many disciplines and data types, from clinical to socioeconomic, are integrated alongside genetic data. The potential of genomics and precision pathogen medicine should not divert away attention from socioeconomic factors that influence infectious disease incidence, such as it is the case of inequality and socioeconomic disparities in tuberculosis. Multidisciplinary studies that encompass the whole spectrum of factors that affect disease are more likely to inform better public health interventions and allocation of resources.

Concluding remarks

During the last decade, pathogen genomics has shown its potential applications outside of research projects and into the clinical laboratory. Recent developments have shown that genetic sequences can be used to identify pathogen species at high accuracy, detect phenotypes of interest such as antimicrobial resistance or virulence, and aid epidemiological studies to inform public health control measures.

However, the seamless implementation of genomics within clinical microbiology is still pending. The work presented in this thesis represents an attempt to expand theoretical concepts in antimicrobial resistance, as well as proposing methodological applications to catalyze the use of pathogen genomics within clinical microbiology

and, eventually, reduce the human burden of infectious diseases.

References

- Adelino TÉR, Giovanetti M, Fonseca V, Xavier J, Abreu ÁS de, Nascimento VA do, Demarchi LHF, Oliveira MAA, Silva VL da, Mello AL de, et al. (2021). Field and classroom initiatives for portable sequence-based monitoring of dengue virus in Brazil. *Nature Communications* 2021 12:1 12(1): 1–12.
- Ahmad S and Mokaddas E (2014). Current status and future trends in the diagnosis and treatment of drug-susceptible and multidrug-resistant tuberculosis. *Journal of Infection and Public Health* 7(2): 75–91.
- Ajawatanawong P, Yanai H, Smittipat N, Disratthakit A, Yamada N, Miyahara R, Nedsuwan S, Imasanguan W, Kantipong P, Chaiyasirinroje B, et al. (2019). A novel Ancestral Beijing sublineage of Mycobacterium tuberculosis suggests the transition site to Modern Beijing sublineages. *Scientific Reports* 9(1): 1–12.
- Alberts B, Heald R, Johnson A, Morgan D, Raff M, Roberts K, Walter P, and Wilson J (2022). *Molecular Biology of the Cell*. Seventh Edition. W. W. Norton & Company.
- Amicone M, Borges V, Alves MJ, Isidro J, Zé-Zé L, Duarte S, Vieira L, Guiomar R, Gomes JP, and Gordo I (2022). Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. *Evolution, Medicine, and Public Health* 10(1): 142–155.
- Andersson DI, Nicoloff H, and Hjort K (2019). Mechanisms and clinical relevance of bacterial heteroresistance. *Nature Reviews Microbiology* 17(8): 479–496.
- Andrews JR, Shah NS, Weissman D, Moll AP, Friedland G, and Gandhi NR (2010). Predictors of Multidrug- and Extensively Drug-Resistant Tuberculosis in a High HIV Prevalence Community. *PLOS ONE* 5(12): e15735.

- Andrews S and Babraham Bioinformatics (2010). FastQC: A quality control tool for high throughput sequence data. *Manual*.
- ARTIC Network (2021). ARTIC nanopore protocol for nCoV2019 novel coronavirus [Internet]. GitHub. (Visited on 01/22/2021).
- Ates LS, Dippenaar A, Ummels R, Piersma SR, Van Der Woude AD, Van Der Kuij K, Le Chevalier F, Mata-Espinosa D, Barrios-Payán J, Marquina-Castillo B, et al. (2018). Mutations in ppe38 block PE-PGRS secretion and increase virulence of *Mycobacterium tuberculosis*. *Nature Microbiology* 3(2): 181–188.
- Attwood SW, Hill SC, Aanensen DM, Connor TR, and Pybus OG (2022). Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nature Reviews Genetics* 2022: 1–16.
- Ayala FJ (1997). Vagaries of the molecular clock. *Proceedings of the National Academy of Sciences* 94(15): 7776–7783.
- Azarian T (2022). Meaningful Use of Pathogen Genomic Data. *mBio* 13(3): e00311–22.
- Bachman MA and LeBar WD (2019). First Principles of Clinical Microbiology: Collection, Handling, and Diagnostics. *Encyclopedia of Microbiology (Fourth Edition)*: 276–281.
- Balloux F, Brønstad Brynildsrud O, Dorp L van, Shaw LP, Chen H, Harris KA, Wang H, and Eldholm V (2018). From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. *Trends in Microbiology* 26(12): 1035–1048.
- Balloux F, Tan C, Swadling L, Richard D, Jenner C, Maini M, and Dorp L van (2022). The past, current and future epidemiological dynamic of SARS-CoV-2. *Oxford Open Immunology* 3(1): iqac003.
- Band VI and Weiss DS (2019). Heteroresistance: A cause of unexplained antibiotic treatment failure? *PLoS Pathogens* 15(6): e1007726.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. (2012). SPAdes: A new genome assem-

- bly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19(5): 455–477.
- Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang ML, Nalla A, Pepper G, Reinhardt A, Xie H, et al. (2020). Cryptic transmission of SARS-CoV-2 in Washington state. *Science* 370(6516): 571–575.
- Behr MA, Edelstein PH, and Ramakrishnan L (2018). Revisiting the timetable of tuberculosis. *The BMJ* 362.
- Bertelli C and Greub G (2013). Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clinical Microbiology and Infection* 19(9): 803–813.
- Bertels F, Gokhale CS, and Traulsen A (2017). Discovering Complete Quasispecies in Bacterial Genomes. *Genetics* 206(4): 2149–2157.
- Berzosa P, De Lucio A, Romay-Barja M, Herrador Z, González V, García L, Fernández-Martínez A, Santana-Morales M, Ncogo P, Valladares B, et al. (2018). Comparison of three diagnostic methods (microscopy, RDT, and PCR) for the detection of malaria parasites in representative samples from Equatorial Guinea 11 Medical and Health Sciences 1108 Medical Microbiology. *Malaria Journal* 17(1): 1–12.
- Biek R, Pybus OG, Lloyd-Smith JO, and Didelot X (2015). Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution* 30(6): 306–313.
- Bille E, Zahar JR, Perrin A, Morelle S, Kriz P, Jolley KA, Maiden MC, Dervin C, Nassif X, and Tinsley CR (2005). A chromosomally integrated bacteriophage in invasive meningococci. *Journal of Experimental Medicine* 201(12): 1905–1913.
- Black A, MacCannell DR, Sibley TR, and Bedford T (2020). Ten recommendations for supporting open pathogen genomic analysis in public health. *Nature Medicine* 26(6): 832–841.
- Bode LGM, Wunnik P van, Vaessen N, Savelkoul PHM, and Smeets LC (2012). Rapid detection of methicillin-resistant *Staphylococcus aureus* in screening samples by relative quantification between the *mecA* gene and the *SA442* gene. *Journal of Microbiological Methods* 89(2): 129–132.

- Bolger AM, Lohse M, and Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15): 2114–2120.
- Borrell S, Teo Y, Giardina F, Streicher EM, Klopper M, Feldmann J, Müller B, Victor TC, and Gagneux S (2013). Epistasis between antibiotic resistance mutations drives the evolution of extensively drug-resistant tuberculosis. *Evolution, Medicine, and Public Health* 2013(1): 65–74.
- Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, and Drummond AJ (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology* 10(4): e1003537.
- Braun KM, Moreno GK, Wagner C, Accola MA, Rehrauer WM, Baker DA, Koelle K, O'Connor DH, Bedford T, Friedrich TC, et al. (2021). Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLOS Pathogens* 17(8): e1009849.
- Britten RJ (1986). Rates of DNA Sequence Evolution Differ Between Taxonomic Groups. *Science* 231(4744): 1393–1398.
- Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR, et al. (2013). Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. *The Lancet* 381(9877): 1551–1560.
- Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, Bohlin J, Alfsnes K, Pettersson JOH, Kirkeleite I, et al. (2018). Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. *Science Advances* 4(10): eaat5869.
- Byrne AW, McEvoy D, Collins AB, Hunt K, Casey M, Barber A, Butler F, Griffin J, Lane EA, McAloon C, et al. (2020). Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ open* 10(8): e039856.
- Cameron DR, Jiang JH, Kostoulas X, Foxwell DJ, and Peleg AY (2016). Vancomycin susceptibility in methicillin-resistant *Staphylococcus aureus* is mediated by YycH

- activation of the WalRK essential two-component regulatory system. *Scientific Reports* 6(1): 30823.
- Campbell F, Strang C, Ferguson N, Cori A, and Jombart T (2018). When are pathogen genome sequences informative of transmission events? *PLoS Pathogens* 14(2): e1006885.
- Carey V, Long L, and Gentleman R (2021). RBGL: An interface to the BOOST graph library.
- Caroff N, Espaze E, Gautreau D, Richet H, and Reynaud A (2000). Analysis of the effects of -42 and -32 ampC promoter mutations in clinical isolates of *Escherichia coli* hyperproducing AmpC. *Journal of Antimicrobial Chemotherapy* 45(6): 783–788.
- Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, et al. (2014). Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature Genetics* 46(3): 279–286.
- Caviedes L, Lee TS, Gilman RH, Sheen P, Spellman E, Lee EH, Berg DE, and Montenegro-James S (2000). Rapid, Efficient Detection and Drug Susceptibility Testing of *Mycobacterium tuberculosis* in Sputum by Microscopic Observation of Broth Cultures. *Journal of Clinical Microbiology* 38(3): 1203–1208.
- Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, Rae D, Grundy S, Turner DJ, Wain J, et al. (2019). Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature Biotechnology* 37(7): 783–792.
- Cintrón M, Hauser JR, Otto C, Persing DH, and Tang YW (2019). Diagnostic Microbiology. *Encyclopedia of Microbiology (Fourth Edition)*: 1–17.
- Clinical and Laboratory Standards Institute (CLSI) (2011). Susceptibility Testing of *Mycobacteria*, *Nocardiae*, and Other Aerobic Actinomycetes; Approved Standard—Second Edition.

- Codjoe FS and Donkor ES (2017). Carbapenem Resistance: A Review. *Medical Sciences* 2018, Vol. 6, Page 1 6(1): 1.
- Cohen KA, Abeel T, Manson McGuire A, Desjardins CA, Munsamy V, Shea TP, Walker BJ, Bantubani N, Almeida DV, Alvarado L, 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 Medicine* 12(9): e1001880.
- Cohen T, Helden PD van, Wilson D, Colijn C, McLaughlin MM, Abubakar I, and Warren RM (2012). Mixed-strain Mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. *Clinical Microbiology Reviews* 25(4): 708–719.
- Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, et al. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* 393(6685): 537–544.
- Colijn C and Gardy J (2014). Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine and Public Health*.
- Coll F, Mallard K, Preston MD, Bentley S, Parkhill J, McNerney R, Martin N, and Clark TG (2012). SpolPred: Rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics*.
- Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, and Clark TG (2014). A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nature Communications* 5(1): 4812.
- Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, Mallard K, Nair M, Miranda A, Alves A, et al. (2015). Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Medicine* 7(1): 51.
- Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, and Gagneux S (2012). Whole-genome sequencing of rifampicin-resistant

- Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics* 44(1): 106–110.
- Cornish-Bowden A (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic acids research* 13(9): 3021–30.
- Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP, and Haydon DT (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences* 275(1637): 887–895.
- COVID-19 Genomics UK (COG-UK) (2020). An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet. Microbe* 1(3): e99–e100.
- Cox HS, Kubica T, Doshetov D, Kebede Y, Rüsch-Gerdess S, and Niemann S (2005). The Beijing genotype and drug resistant tuberculosis in the Aral Sea region of Central Asia. *Respiratory Research* 6(1): 134.
- Cudini J, Roy S, Houldcroft CJ, Bryant JM, Depledge DP, Tutill H, Veys P, Williams R, Worth AJJ, Tamuri AU, et al. (2019). Human cytomegalovirus haplotype reconstruction reveals high diversity due to superinfection and evidence of within-host recombination. *Proceedings of the National Academy of Sciences* 116(12): 5693–5698.
- Cuevas JM, Geller R, Garijo R, López-Aldeguer J, and Sanjuán R (2015). Extremely High Mutation Rate of HIV-1 In Vivo. *PLOS Biology* 13(9): e1002251.
- D’Costa VM, McGrann KM, Hughes DW, and Wright GD (2006). Sampling the Antibiotic Resistome. *Science* 311(5759): 374–377.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27(15): 2156–2158.
- Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, Pearson CAB, Russell TW, Tully DC, Washburne AD, et al. (2021a). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science (New York, N.Y.)*

- Davies NG, Jarvis CI, Zandvoort K van, Clifford S, Sun FY, Funk S, Medley G, Jafari Y, Meakin SR, Lowe R, et al. (2021b). Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature*: 1–5.
- De Maio N, Walker C, Borges R, Weilguny L, Slodkowitz G, and Goldman N (2020). Masking strategies for SARS-CoV-2 alignments. *Virological*.
- De Maio N, Worby CJ, Wilson DJ, and Stoesser N (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLOS Computational Biology* 14(4): e1006117.
- Dellicour S, Baele G, Dudas G, Faria NR, Pybus OG, Suchard MA, Rambaut A, and Lemey P (2018). Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nature Communications* 9(1): 2222.
- Deng X, Gu W, Federman S, Plessis L du, Pybus OG, Faria NR, Wang C, Yu G, Bushnell B, Pan CY, et al. (2020). Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* 369(6503): 582–587.
- Didelot X, Bowden R, Wilson DJ, Peto TEA, and Crook DW (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics* 13(9): 601–612.
- Didelot X, Croucher NJ, Bentley SD, Harris SR, and Wilson DJ (2018). Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Research* 46(22): e134–e134.
- Didelot X, Fraser C, Gardy J, and Colijn C (2017). Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks. *Molecular Biology and Evolution* 34(4): 997–1007.
- Didelot X, Gardy J, and Colijn C (2014). Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data. *Molecular Biology and Evolution* 31(7): 1869–1879.
- Didelot X, Walker AS, Peto TE, Crook DW, and Wilson DJ (2016). Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology*.

- Dinis JM, Florek NW, Fatola OO, Moncla LH, Mutschler JP, Charlier OK, Meece JK, Belongia EA, and Friedrich TC (2016). Deep Sequencing Reveals Potential Antigenic Variants at Low Frequencies in Influenza A Virus-Infected Humans. *Journal of Virology* 90(7): 3355–3365.
- Domingo E, Menéndez-Arias L, Quiñones-Mateu ME, Holguín A, Gutiérrez-Rivas M, Martínez MA, Quer J, Novella IS, and Holland JJ (1997). Viral quasispecies and the problem of vaccine-escape and drug-resistant mutants. *Progress in Drug Research. Fortschritte Der Arzneimittelforschung. Progres Des Recherches Pharmaceutiques* 48: 99–128.
- Domingo E (1998). Quasispecies and the implications for virus persistence and escape. *Clinical and Diagnostic Virology* 10(2): 97–101.
- Domingo E, Sheldon J, and Perales C (2012). Viral Quasispecies Evolution. *Microbiology and Molecular Biology Reviews* 76(2): 159–216.
- Dordel J, Kim C, Chung M, Pardos de la Gándara M, Holden MTJ, Parkhill J, Lencastre H de, Bentley SD, and Tomasz A (2014). Novel determinants of antibiotic resistance: identification of mutated loci in highly methicillin-resistant subpopulations of methicillin-resistant *Staphylococcus aureus*. *mBio* 5(2): e01000.
- Dorp L van, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CC, Boshier FA, et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution*.
- Drobniewski F (2005). Drug-Resistant Tuberculosis, Clinical Virulence, and the Dominance of the Beijing Strain Family in Russia. *JAMA* 293(22): 2726.
- Drummond AJ, Ho SY, Phillips MJ, and Rambaut A (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*.
- Drummond AJ, Pybus OG, Rambaut A, Forsberg R, and Rodrigo AG (2003). Measurably evolving populations. *Trends in Ecology & Evolution* 18(9): 481–488.
- Duchêne S, Holt KE, Weill FX, Le Hello S, Hawkey J, Edwards DJ, Fourment M, and Holmes EC (2016). Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics* 2(11): e000094.

- Duchêne S, Duchêne D, Holmes EC, and Ho SY (2015). The Performance of the Date-Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data. *Molecular Biology and Evolution* 32(7): 1895–1906.
- Duchêne S and Holmes EC (2018). Estimating evolutionary rates in giant viruses using ancient genomes. *Virus Evolution* 4(1): vey006.
- Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, Park DJ, Ladner JT, Arias A, Asogun D, et al. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 544(7650): 309–315.
- Eigen M and Schuster P (1977). A principle of natural self-organization: Part A: Emergence of the hypercycle. *Naturwissenschaften* 64(11): 541–565.
- Elbe S and Buckland-Merrett G (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* 1(1): 33–46.
- Eldholm V and Balloux F (2016). Antimicrobial Resistance in Mycobacterium tuberculosis : The Odd One Out. *Trends in Microbiology* 24(8): 637–648.
- Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, and Balloux F (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature Communications* 6(1): 7119.
- Eldholm V, Rieux A, Monteserin J, Lopez JM, Palmero D, Lopez B, Ritacco V, Dide-
lot X, and Balloux F (2016). Impact of HIV co-infection on the evolution and transmission of multidrug-resistant tuberculosis. *eLife* 5 (AUGUST).
- Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, Grundman H, Hasman H, Holden MTG, Hopkins KL, et al. (2017). The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clinical Microbiology and Infection* 23(1): 2–22.
- Elliott P, Haw D, Wang H, Eales O, Walters CE, Ainslie KEC, Atchison C, Fronterre C, Diggle PJ, Page AJ, et al. (2021). Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the Delta variant. *Science* 374(6574): eabl9551.

- Engelthaler DM, Streicher EM, Kelley EJ, Allender CJ, Wiggins K, Jimenez D, Lemmer D, Vittinghoff E, Theron G, Sirgel FA, et al. (2019). Minority Mycobacterium tuberculosis Genotypic Populations as an Indicator of Subsequent Phenotypic Resistance. *American Journal of Respiratory Cell and Molecular Biology* 61(6): 789–791.
- Ershova JV, Kurbatova EV, Moonan PK, and Cegielski JP (2014). Mortality Among Tuberculosis Patients With Acquired Resistance to Second-line Antituberculosis Drugs—United States, 1993–2008. *Clinical Infectious Diseases* 59(4): 465–472.
- Evans D, Cowen S, Kammel M, O'Sullivan DM, Stewart G, Grunert HP, Moran-Gilad J, Verwilt J, In J, Vandesompele J, et al. (2021). The Dangers of Using Cq to Quantify Nucleic Acid in Biological Samples: A Lesson From COVID-19. *Clinical chemistry* 68(1): 153–162.
- Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, and Foll M (2013). Robust Demographic Inference from Genomic and SNP Data. *PLOS Genetics* 9(10): e1003905.
- Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L, et al. (2012). A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance. *BMJ Open* 2(3): e001124.
- Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, et al. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nature genetics* 45(10): 1183–9.
- Faria NR, Quick J, Claro IM, Théze J, De Jesus JG, Giovanetti M, Kraemer MU, Hill SC, Black A, Da Costa AC, et al. (2017). Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* 2017 546:7658 546(7658): 406–410.
- Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pèpin J, et al. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346(6205): 56–61.

- Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, Vogels CBF, Brito AF, Alpert T, Muyombwe A, et al. (2020). Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* 181(5): 990–996.e5.
- Fishbein S, Wyk N van, Warren RM, and Sampson SL (2015). Phylogeny to function: PE/PPE protein evolution and impact on Mycobacterium tuberculosis pathogenicity. *Molecular Microbiology* 96(5): 901–916.
- Florensa AF, Kaas RS, Clausen PTL, Aytan-Aktug D, and Aarestrup FM (2022). ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microbial Genomics* 8(1).
- Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, and Fortune SM (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature Genetics* 45(7): 784–790.
- Forrellad MA, Klepp LI, Gioffré A, García JS, Morbidoni HR, Paz Santangelo M de la, Cataldi AA, and Bigi F (2013). Virulence factors of the mycobacterium tuberculosis complex. *Virulence* 4(1): 3–66.
- Fraser C, Lythgoe K, Leventhal GE, Shirreff G, Hollingsworth TD, Alizon S, and Bonhoeffer S (2014). Virulence and pathogenesis of HIV-1 infection: An evolutionary perspective. *Science* 343(6177).
- Frise R, Bradley K, Doremalen N van, Galiano M, Elderfield RA, Stilwell P, Ashcroft JW, Fernandez-Alonso M, Miah S, Lackenby A, et al. (2016). Contact transmission of influenza virus between ferrets imposes a looser bottleneck than respiratory droplet transmission allowing propagation of antiviral resistance. *Scientific Reports* 6: 29793.
- Gagneux S (2018). Ecology and evolution of Mycobacterium tuberculosis. *Nature Reviews Microbiology*.

- Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, and Bohannon BJ (2006). The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis*. *Science* 312(5782): 1944–1946.
- Galvani AP (2003). Epidemiology meets evolutionary ecology. *Trends in Ecology & Evolution* 18(3): 132–139.
- Gandhi NR, Nunn P, Dheda K, Schaaf HS, Zignol M, Soolingen D van, Jensen P, and Bayona J (2010). Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *The Lancet* 375(9728): 1830–1843.
- Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodtkin E, Rempel S, Moore R, Zhao Y, Holt R, et al. (2011). Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *New England Journal of Medicine*.
- Gardy JL and Loman NJ (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics* 19: 9–20.
- Genton B, D’Acremont V, Rare L, Baea K, Reeder JC, Alpers MP, and Müller I (2008). *Plasmodium vivax* and Mixed Infections Are Associated with Severe Malaria in Children: A Prospective Cohort Study from Papua New Guinea. *PLOS Medicine* 5(6): e127.
- Giandhari J, Pillay S, Wilkinson E, Tegally H, Sinayskiy I, Schuld M, Lourenço J, Chimukangara B, Lessells R, Moosa Y, et al. (2021). Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. *International Journal of Infectious Diseases* 103: 234–241.
- Giani AM, Gallo GR, Gianfranceschi L, and Formenti G (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal* 18: 9–19.
- Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, et al. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*.

- Glynn JR, Whiteley J, Bifani PJ, Kremer K, and Soolingen D van (2002). Worldwide Occurrence of Beijing/W Strains of *Mycobacterium tuberculosis* : A Systematic Review. *Emerging Infectious Diseases* 8(8): 843–849.
- Gonzales MJ (1989). Chinese Plantation Workers and Social Conflict in Peru in the late Nineteenth Century. *Journal of Latin American Studies* 21(3): 385–424.
- Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, Fabre S, Kleiner G, Polanco J, Khan Z, et al. (2020). Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* 369(6501): 297–301.
- Grandjean L, Gilman RH, Iwamoto T, Köser CU, Coronel J, Zimic M, Török ME, Ayabina D, Kendall M, Fraser C, et al. (2017). Convergent evolution and topologically disruptive polymorphisms among multidrug-resistant tuberculosis in Peru. *PLOS ONE* 12(12): e0189838.
- Grandjean L, Gilman RH, Martin L, Soto E, Castro B, Lopez S, Coronel J, Castillo E, Alarcon V, Lopez V, et al. (2015). Transmission of Multidrug-Resistant and Drug-Susceptible Tuberculosis within Households: A Prospective Cohort Study. *PLOS Medicine* 12(6): e1001843.
- Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, and Holmes EC (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science* 303(5656): 327–332.
- Greub G, Kebbi-Beghdadi C, Bertelli C, Collyn F, Riederer BM, Yersin C, Croxatto A, and Raoult D (2009). High throughput sequencing and proteomics to identify immunogenic proteins of a new pathogen: the dirty genome approach. *PloS One* 4(12): e8423.
- Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, Tan AL, Paul LM, Brackney DE, Grewal S, et al. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology* 20(1): 8.

- Gupta B, Gupta P, Sharma A, Singh V, Dash AP, and Das A (2010). High proportion of mixed-species Plasmodium infections in India revealed by PCR diagnostic assay. *Tropical Medicine & International Health* 15(7): 819–824.
- Gygli SM, Borrell S, Trauner A, and Gagneux S (2017). Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives. *FEMS Microbiology Reviews* 41(3): 354–373.
- Hahne F and Ivanek R (2016). Visualizing Genomic Data Using Gviz and Bioconductor. *Statistical Genomics: Methods and Protocols*: 335–351.
- El-Halfawy OM and Valvano MA (2015). Antimicrobial Heteroresistance: an Emerging Field in Need of Clarity. *Clinical Microbiology Reviews* 28(1): 191–207.
- Hall M, Woolhouse M, and Rambaut A (2015). Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLOS Computational Biology* 11(12): e1004613.
- Han A, Czajkowski LM, Donaldson A, Baus HA, Reed SM, Athota RS, Bristol T, Rosas LA, Cervantes-Medina A, Taubenberger JK, et al. (2019). A Dose-finding Study of a Wild-type Influenza A(H3N2) Virus in a Healthy Volunteer Human Challenge Model. *Clinical Infectious Diseases* 69(12): 2082–2090.
- Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J, et al. (2013). Whole-genome sequencing for analysis of an outbreak of methicillin-resistant Staphylococcus aureus: A descriptive study. *The Lancet Infectious Diseases*.
- Hart WS, Miller E, Andrews NJ, Waight P, Maini PK, Funk S, and Thompson RN (2022). Generation time of the alpha and delta SARS-CoV-2 variants: an epidemiological analysis. *The Lancet Infectious Diseases* 22(5): 603–610.
- Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, Ludden C, Reeve R, Rambaut A, Peacock SJ, et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* 19(7): 409–424.
- Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, and Aarestrup FM (2014). Rapid Whole-Genome Sequencing for Detection and

- Characterization of Microorganisms Directly from Clinical Samples. *Journal of Clinical Microbiology* 52(1): 139–146.
- Heijden YF van der, Karim F, Mufamadi G, Zako L, Chinappa T, Shepherd BE, Maruri F, Moosa MYS, Sterling TR, and Pym AS (2017). Isoniazid-monoresistant tuberculosis is associated with poor treatment outcomes in Durban, South Africa. *The International Journal of Tuberculosis and Lung Disease* 21(6): 670–676.
- Helb D, Jones M, Story E, Boehme C, Wallace E, Ho K, Kop J, Owens MR, Rodgers R, Banada P, et al. (2010). Rapid detection of *Mycobacterium tuberculosis* and rifampin resistance by use of on-demand, near-patient technology. *Journal of clinical microbiology* 48(1): 229–237.
- Hicks ND, Yang J, Zhang X, Zhao B, Grad YH, Liu L, Ou X, Chang Z, Xia H, Zhou Y, et al. (2018). Clinically prevalent mutations in *Mycobacterium tuberculosis* alter propionate metabolism and mediate multidrug tolerance. *Nature microbiology* 3(9): 1032–1042.
- Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, and Cooper A (2011). Time-dependent rates of molecular evolution. *Molecular Ecology* 20(15): 3087–3101.
- Horn DL, Hewlett D, Haas WH, Butler WR, Alfalla C, Tan E, Levine A, Nayak A, and Opal SM (1994). Superinfection with rifampin-isoniazid-streptomycin-ethambutol (RISE)-resistant tuberculosis in three patients with AIDS: Confirmation by polymerase chain reaction fingerprinting. *Annals of Internal Medicine* 121(2): 115–116.
- Hou TY, Chiang-Ni C, and Teng SH (2019). Current status of MALDI-TOF mass spectrometry in clinical microbiology. *Journal of Food and Drug Analysis* 27(2): 404–414.
- Houlihan CF, Frampton D, Bridget Ferns R, Raffle J, Grant P, Reidy M, Hail L, Thomson K, Mattes F, Kozlakidis Z, et al. (2018). Use of whole-genome sequencing in the investigation of a nosocomial influenza virus outbreak. *Journal of Infectious Diseases*.

- Huang W, Li L, Myers JR, and Marth GT (2012). ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)* 28(4): 593–4.
- Ismail SA, Saliba V, Lopez Bernal J, Ramsay ME, and Ladhani SN (2021). SARS-CoV-2 infection and transmission in educational settings: a prospective, cross-sectional analysis of infection clusters and outbreaks in England. *The Lancet Infectious Diseases* 21(3): 344–353.
- IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) (1984). Nomenclature and symbolism for amino acids and peptides. Recommendations 1983. *European journal of biochemistry* 138(1): 9–37.
- Jangir PK, Yang Q, Shaw LP, Caballero JD, Ogunlana L, Wheatley R, Walsh T, and MacLean RC (2022). Pre-existing chromosomal polymorphisms in pathogenic *E. coli* potentiate the evolution of resistance to a last-resort antibiotic. *eLife* 11: e78834.
- Jenkins GM, Rambaut A, Pybus OG, and Holmes EC (2002). Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis. *Journal of Molecular Evolution* 54(2): 156–165.
- Jolivet-Gougeon A, Kovacs B, Le Gall-David S, Le Bars H, Bousarghin L, Bonnaure-Mallet M, Lobel B, Guillé F, Soussy CJ, and Tenke P2 (2011). Bacterial hypermutation: clinical implications. *Journal of Medical Microbiology* 60(5): 563–573.
- Jombart T, Eggo RM, Dodd PJ, and Balloux F (2011). Reconstructing disease outbreaks from genetic data: A graph approach. *Heredity*.
- Kaplan G, Post FA, Moreira AL, Wainwright H, Kreiswirth BN, Tanverdi M, Mathema B, Ramaswamy SV, Walther G, Steyn LM, et al. (2003). Mycobacterium tuberculosis Growth at the Cavity Surface: A Microenvironment with Failed Immunity. *Infection and Immunity* 71(12): 7099–7108.
- Kapopoulou A, Lew JM, and Cole ST (2011). The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis* 91(1): 8–13.

- Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, et al. (2011). *Helicobacter pylori* genome evolution during human infection. *Proceedings of the National Academy of Sciences* 108(12): 5033–5038.
- Keshavjee S and Farmer PE (2012). Tuberculosis, Drug Resistance, and the History of Modern Medicine. *New England Journal of Medicine* 367(10): 931–936.
- Kim DW and Cha CJ (2021). Antibiotic resistance from the One-Health perspective: understanding and controlling antimicrobial resistance transmission. *Experimental & Molecular Medicine* 53(3): 301–309.
- Kimura M (1987). Molecular evolutionary clock and the neutral theory. *Journal of Molecular Evolution* 26(1): 24–33.
- Klompas M, Baker MA, Rhee C, Tucker R, Fiumara K, Griesbach D, Bennett-Rizzo C, Salmasian H, Wang R, Wheeler N, et al. (2021). A SARS-CoV-2 cluster in an acute care hospital. *Annals of Internal Medicine* 174(6): 794–802.
- Knaus BJ and Grünwald NJ (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources* 17(1): 44–53.
- Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, et al. (2012). Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology. *PLOS Pathogens* 8(8): e1002824.
- Kotepui M, Kotepui KU, De Jesus Milanez G, and Masangkay FR (2020a). Plasmodium spp. mixed infection leading to severe malaria: a systematic review and meta-analysis. *Scientific Reports* 2020 10:1 10(1): 1–12.
- (2020b). Summary of discordant results between rapid diagnosis tests, microscopy, and polymerase chain reaction for detecting Plasmodium mixed infection: a systematic review and meta-analysis. *Scientific Reports* 2020 10:1 10(1): 1–17.
- Kozlov AM, Darriba D, Flouri T, Morel B, and Stamatakis A (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35(21): 4453–4455.

- Kuhner MK and Felsenstein J (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* 11(3): 459–468.
- Kwong JC, Mccallum N, Sintchenko V, and Howden BP (2015). Whole genome sequencing in clinical and public health microbiology. *Pathology* 47: 199–210.
- Ladhani SN, Chow JY, Janarthanan R, Fok J, Crawley-Boevey E, Vusirikala A, Fernandez E, Perez MS, Tang S, Dun-Campbell K, et al. (2020). Increased risk of SARS-CoV-2 infection in staff working across different care homes: enhanced COVID-19 outbreak investigations in London care Homes. *Journal of Infection*.
- Lee LYW, Rozmanowski S, Pang M, Charlett A, Anderson C, Hughes GJ, Barnard M, Peto L, Vipond R, Sienkiewicz A, et al. (2022). Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infectivity by Viral Load, S Gene Variants and Demographic Factors, and the Utility of Lateral Flow Devices to Prevent Transmission. *Clinical Infectious Diseases* 74(3): 407–415.
- Lee RS and Behr MA (2016). Does choice matter? reference-based alignment for molecular epidemiology of tuberculosis. *Journal of Clinical Microbiology* 54(7): 1891–1895.
- Leitner T (2019). Phylogenetics in HIV transmission: Taking within-host diversity into account. *Current Opinion in HIV and AIDS* 14(3): 181–187.
- Lemieux JE, Siddle KJ, Shaw BM, Loreth C, Schaffner SF, Gladden-Young A, Adams G, Fink T, Tomkins-Tinch CH, Krasilnikova LA, et al. (2021). Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* 371(6529): eabe3261.
- Li H (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18): 3094–3100.
- Li H and Durbin R (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26(5): 589–595.

- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16): 2078–2079.
- Lieberman TD, Wilson D, Misra R, Xiong LL, Moodley P, Cohen T, and Kishony R (2016). Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis. *Nature medicine* 22(12): 1470–1474.
- Lowen AC, Mubareka S, Steel J, and Palese P (2007). Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature. *PLOS Pathogens* 3(10): e151.
- Lythgoe KA, Hall M, Ferretti L, Cesare Md, MacIntyre-Cockett G, Trebes A, Andersson M, Otecko N, Wise EL, Moore N, et al. (2020). Shared SARS-CoV-2 diversity suggests localised transmission of minority variants. *bioRxiv*.
- Mahé P, Arsac M, Chatellier S, Monnin V, Perrot N, Mailler S, Girard V, Ramjeet M, Surre J, Lacroix B, et al. (2014). Automatic identification of mixed bacterial species fingerprints in a MALDI-TOF mass-spectrum. *Bioinformatics* 30(9): 1280–1286.
- Maltha J, Gillet P, Cnops L, Van Den Ende J, Van Esbroeck M, and Jacobs J (2010). Malaria rapid diagnostic tests: Plasmodium falciparum infections with high parasite densities may generate false positive Plasmodium vivax pLDH lines. *Malaria Journal* 9(1): 1–7.
- Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, Barry CE, Brand J, Chapman SB, Cho SN, Gabrielian A, et al. (2017). Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nature Genetics* 49(3): 395–402.
- Mascellino MT, Porowska B, De Angelis M, and Oliva A (2017). Antibiotic susceptibility, heteroresistance, and updated treatment strategies in Helicobacter pylori infection. *Drug Design, Development and Therapy* 11: 2209–2220.
- Matsuhashi M, Song MD, Ishino F, Wachi M, Doi M, Inoue M, Ubukata K, Yamashita N, and Konno M (1986). Molecular cloning of the gene of a penicillin-binding protein

- supposed to cause high resistance to beta-lactam antibiotics in *Staphylococcus aureus*. *Journal of Bacteriology* 167(3): 975–980.
- McCrone JT and Lauring AS (2016). Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling. *Journal of Virology* 90(15): 6884–6895.
- McEvoy CR, Cloete R, Müller B, Schürch AC, Helden PD van, Gagneux S, Warren RM, and Gey van Pittius NC (2012). Comparative Analysis of *Mycobacterium tuberculosis* *pe* and *ppe* Genes Reveals High Sequence Variation and an Apparent Absence of Selective Constraints. *PLOS ONE* 7(4): e30593.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20(9): 1297–1303.
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, and Cunningham F (2016). The Ensembl Variant Effect Predictor. *Genome Biology* 2016 17:1 17(1): 1–14.
- Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, Farhat MR, Guthrie JL, Laukens K, Miotto P, et al. (2019). Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nature Reviews Microbiology* 17(9): 533–545.
- Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, et al. (2011). Prospective Genomic Characterization of the German Enterohemorrhagic *Escherichia coli* O104:H4 Outbreak by Rapid Next Generation Sequencing Technology. *PLoS ONE* 6(7): e22751.
- Menardo F, Duchêne S, Brites D, and Gagneux S (2019). The molecular clock of *Mycobacterium tuberculosis*. *PLOS Pathogens* 15(9): e1008067.
- Meng B, Kemp SA, Papa G, Datir R, Ferreira IATM, Marelli S, Harvey WT, Lytras S, Mohamed A, Gallo G, et al. (2021). Recurrent emergence of SARS-CoV-2 spike

- deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Reports* 35(13): 109292.
- Meng XL (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics* 12(2): 685–726.
- Menzies D, Benedetti A, Paydar A, Martin I, Royce S, Pai M, Vernon A, Lienhardt C, and Burman W (2009). Effect of Duration and Intermittency of Rifampin on Tuberculosis Treatment Outcomes: A Systematic Review and Meta-Analysis. *PLoS Medicine* 6(9): e1000146.
- Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, Curran MD, Parmar S, Caller LG, Caddy SL, et al. (2020). Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *The Lancet Infectious Diseases*.
- Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, West K, Qu J, Baniecki ML, Gladden-Young A, et al. (2017). Zika virus evolution and spread in the Americas. *Nature*.
- Mohanty S, Dal Molin M, Ganguli G, Padhi A, Jena P, Selchow P, Sengupta S, Meuli M, Sander P, and Sonawane A (2016). Mycobacterium tuberculosis EsxO (Rv2346c) promotes bacillary survival by inducing oxidative stress mediated genomic instability in macrophages. *Tuberculosis* 96: 44–57.
- Mongkolrattanothai K, Gray BM, Mankin P, Stanfill AB, Pearl RH, Wallace LJ, and Ve-gunta RK (2011). Simultaneous carriage of multiple genotypes of Staphylococcus aureus in children. *Journal of Medical Microbiology* 60(3): 317–322.
- Moore DA, Evans CA, Gilman RH, Caviedes L, Coronel J, Vivar A, Sanchez E, Piñedo Y, Saravia JC, Salazar C, et al. (2006). Microscopic-Observation Drug-Susceptibility Assay for the Diagnosis of TB. *New England Journal of Medicine* 355(15): 1539–1550.

- Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, and Soubeyrand S (2012). A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. *PLoS Computational Biology*.
- Murphy BR, Clements ML, Madore HP, Steinberg J, O'Donnell S, Betts R, Demico D, Reichman RC, Dolin R, and Maassab HF (1984). Dose Response of Cold-Adapted, Reassortant Influenza A/California/10/78 Virus (H1N1) in Adult Volunteers. *Journal of Infectious Diseases* 149(5): 816–816.
- Murray GGR, Wang F, Harrison EM, Paterson GK, Mather AE, Harris SR, Holmes MA, Rambaut A, and Welch JJ (2016). The effect of genetic structure on molecular dating and tests for temporal signal. *Methods in Ecology and Evolution* 7(1): 80–89.
- Murray PR, Rosenthal KS, and Pfaller MA (2021). Molecular Diagnosis. *Medical Microbiology*. 9th ed.: 24–29.
- Mwangi MM, Wu SW, Zhou Y, Sieradzki K, Lencastre H de, Richardson P, Bruce D, Rubin E, Myers E, Siggia ED, et al. (2007). Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. *Proceedings of the National Academy of Sciences* 104(22): 9451–9456.
- Nardell E, McInnis B, Thomas B, and Weidhaas S (1986). Exogenous Reinfection with Tuberculosis in a Shelter for the Homeless. *New England Journal of Medicine* 315(25): 1570–1575.
- Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfidia O, Antoine R, Niyigena EB, Mulders W, Fissette K, et al. (2020). A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region. *Nature Communications* 11(1): 2917.
- Nguyen TNA, Anton-Le Berre V, Bañuls AL, and Nguyen TVA (2019). Molecular Diagnosis of Drug-Resistant Tuberculosis; A Literature Review. *Frontiers in Microbiology* 10 (MAR): 794.
- Nicoloff H, Hjort K, Levin BR, and Andersson DI (2019). The high prevalence of antibiotic heteroresistance in pathogenic bacteria is mainly caused by gene amplification. *Nature Microbiology* 4(3): 504–514.

- Niemann S, Richter E, Rüsç-Gerdes S, Schlaak M, and Greinert U (2000). Double infection with a resistant and a multidrug-resistant strain of *Mycobacterium tuberculosis*. *Emerging Infectious Diseases* 6(5).
- Nimmo C, Shaw LP, Doyle R, Williams R, Brien K, Burgess C, Breuer J, Balloux F, and Pym AS (2019). Whole genome sequencing *Mycobacterium tuberculosis* directly from sputum identifies more genetic diversity than sequencing from culture. *BMC Genomics* 20(1): 389.
- Nsofor CA, Jiang Q, Wu J, Gan M, Liu Q, Zuo T, Zhu G, and Gao Q (2017). Transmission is a Noticeable Cause of Resistance Among Treated Tuberculosis Patients in Shanghai, China. *Scientific Reports* 7(1): 7691.
- Nundu SS, Culleton R, Simpson SV, Arima H, Muyembe JJ, Mita T, Ahuka S, and Yamamoto T (2021). Malaria parasite species composition of *Plasmodium* infections among asymptomatic and symptomatic school-age children in rural and urban areas of Kinshasa, Democratic Republic of Congo. *Malaria Journal* 20(1): 1–13.
- Odone A, Calderon R, Becerra MC, Zhang Z, Contreras CC, Yataco R, Galea J, Lecca L, Bonds MH, Mitnick CD, et al. (2016). Acquired and Transmitted Multidrug Resistant Tuberculosis: The Role of Social Determinants. *PLOS ONE* 11(1): e0146642.
- Pang Y, Zhou Y, Zhao B, Liu G, Jiang G, Xia H, Song Y, Shang Y, Wang S, and Zhao YL (2012). Spoligotyping and Drug Resistance Analysis of *Mycobacterium tuberculosis* Strains from National Survey in China. *PLoS ONE* 7(3): e32976.
- Paradis E, Claude J, and Strimmer K (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20(2): 289–290.
- Paulson T (2013). Epidemiology: A mortal foe. *Nature* 502(7470): S2–S3.
- Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, and Loose M (2021). Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology* 39(4): 442–450.
- Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, Gey van Pittius NC, Glynn JR, Crampin AC, Alves A, et al. (2016). Recombination in *pe/ppe* genes

- contributes to genetic variation in *Mycobacterium tuberculosis* lineages. *BMC Genomics* 17(1): 151.
- Planas D, Veyer D, Baidaliuk A, Planchais C, Porrot F, Robillard N, Puech J, Prot M, Gallais F, Gantner P, et al. (2021). Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature*.
- Popa A, Genger JW, Nicholson MD, Penz T, Schmid D, Aberle SW, Agerer B, Lercher A, Endler L, Colaço H, et al. (2020). Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Science Translational Medicine* 12(573): 2555.
- Pybus OG and Rambaut A (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics* 2009 10:8 10(8): 540–550.
- Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 2016 530:7589 530(7589): 228–232.
- R Core Team and R Foundation for Statistical Computing (2021). R: A Language and Environment for Statistical Computing.
- Rambaut A, Holmes EC, Hill V, O'Toole Á, McCrone JT, Ruis C, Plessis L du, and Pybus OG (2020). A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv*: 2020.04.17.046086.
- Rambaut A, Posada D, Crandall KA, and Holmes EC (2004). The causes and consequences of HIV evolution. *Nature Reviews Genetics* 5(1): 52–61.
- Rengarajan J, Bloom BR, and Rubin EJ (2005). From The Cover: Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proceedings of the National Academy of Sciences* 102(23): 8327–8332.
- Revell LJ (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3(2): 217–223.
- Rie A van, Victor TC, Richardson M, Johnson R, Spuy GD van der, Murray EJ, Beyers N, Pittius NCG van, Helden PD van, and Warren RM (2005). Reinfection and Mixed Infection Cause Changing *Mycobacterium tuberculosis* Drug-Resistance

- Patterns. *American Journal of Respiratory and Critical Care Medicine* 172(5): 636–642.
- Rieux A and Balloux F (2016). Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular Ecology* 25(9): 1911–1924.
- Ritacco V, López B, Cafrune PI, Ferrazoli L, Suffys PN, Candia N, Vásquez L, Realpe T, Fernández J, Lima KV, et al. (2008). Mycobacterium tuberculosis strains of the Beijing genotype are rarely observed in tuberculosis patients in South America. *Memorias do Instituto Oswaldo Cruz* 103(5): 489–492.
- Robins-Browne RM, Holt KE, Ingle DJ, Hocking DM, Yang J, and Tauschek M (2016). Are Escherichia coli Pathotypes Still Relevant in the Era of Whole-Genome Sequencing? *Frontiers in Cellular and Infection Microbiology* 6.
- Robinson D and Foulds L (1981). Comparison of phylogenetic trees. *Mathematical Biosciences* 53(1): 131–147.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, and Mesirov JP (2011). Integrative genomics viewer. *Nature Biotechnology* 29(1): 24–26.
- San JE, Ngcapu S, Kanzi AM, Tegally H, Fonseca V, Giandhari J, Wilkinson E, Nelson CW, Smidt W, Kiran AM, et al. (2021). Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evolution* 7(1): 41.
- Sanger F, Nicklen S, and Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* 74(12): 5463–5467.
- Sanjuán R (2012). From Molecular Genetics to Phylodynamics: Evolutionary Relevance of Mutation Rates Across Viruses. *PLOS Pathogens* 8(5): e1002685.
- Satta G, Lipman M, Smith GP, Arnold C, Kon OM, and McHugh TD (2018). Mycobacterium tuberculosis and whole-genome sequencing: how close are we to unleashing its full potential? *Clinical Microbiology and Infection* 24(6): 604–609.
- Saunders NJ, Trivedi UH, Thomson ML, Doig C, Laurenson IF, and Blaxter ML (2011). Deep resequencing of serial sputum isolates of Mycobacterium tuberculosis during

- therapeutic failure due to poor compliance reveals stepwise mutation of key resistance genes on an otherwise stable genetic background. *Journal of Infection* 62(3): 212–217.
- Schliep KP (2011). phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4): 592–593.
- Sender R, Bar-On YM, Gleizer S, Bernshtein B, Flamholz A, Phillips R, and Milo R (2021). The total number and mass of SARS-CoV-2 virions. *Proceedings of the National Academy of Sciences of the United States of America* 118(25).
- Shafer R, Singh S, Larkin C, and Small P (1995). Exogenous reinfection with multidrug-resistant *Mycobacterium tuberculosis* in an immunocompetent patient. *Tubercle and Lung Disease* 76(6): 575–577.
- Shamputa IC, Jugheli L, Sadradze N, Willery E, Portaels F, Supply P, and Rigouts L (2006). Mixed infection and clonal representativeness of a single sputum sample in tuberculosis patients from a penitentiary hospital in Georgia. *Respiratory research* 7(1).
- Sherman DR, Mdluli K, Hickey MJ, Arain TM, Morris SL, Barry CE, and Stover CK (1996). Compensatory *ahpC* Gene Expression in Isoniazid-Resistant *Mycobacterium tuberculosis*. *Science* 272(5268): 1641–1643.
- Shrestha A, Picoy J, Torres Ortiz A, Moore DA, Gilman RH, Coronel J, and Grandjean L (2021). A case report of transmission and disease caused by *Mycobacterium caprae* and *Mycobacterium bovis* in Lima, Peru. *BMC Infectious Diseases* 21(1): 1–6.
- Sikkema RS, Pas SD, Nieuwenhuijse DF, O'Toole Á, Verweij JJ, Linden A van der, Chestakova I, Schapendonk C, Pronk M, Lexmond P, et al. (2020). COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. *The Lancet Infectious Diseases*.
- Silva Filipe A da, Shepherd JG, Williams T, Hughes J, Aranday-Cortes E, Asamaphan P, Ashraf S, Balcazar C, Bruncker K, Campbell A, et al. (2021). Genomic epidemiology

- reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. *Nature Microbiology* 6(1): 112–122.
- Simmonds P (2004). Genetic diversity and evolution of hepatitis C virus – 15 years on. *Journal of General Virology* 85(11): 3173–3188.
- Sitali L, Chipeta J, Miller JM, Moonga HB, Kumar N, Moss WJ, and Michelo C (2015). Patterns of mixed Plasmodium species infections among children six years and under in selected malaria hyper-endemic communities of Zambia: Population-based survey observations. *BMC Infectious Diseases* 15(1): 1–6.
- Small PM, Shafer RW, Hopewell PC, Singh SP, Murphy MJ, Desmond E, Sierra MF, and Schoolnik GK (1993). Exogenous Reinfection with Multidrug-Resistant Mycobacterium tuberculosis in Patients with Advanced HIV Infection. *New England Journal of Medicine* 328(16).
- Smith G (2020). The paradox of big data. *SN Applied Sciences* 2(6): 1041.
- Smith G and Cordes J (2019). The 9 Pitfalls of Data Science. Oxford University Press.
- Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, and Segre JA (2012). Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Science Translational Medicine* 4(148).
- Spinelli MA, Glidden DV, Gennatas ED, Bielecki M, Beyrer C, Rutherford G, Chambers H, Goosby E, and Gandhi M (2021). Importance of non-pharmaceutical interventions in lowering the viral inoculum to reduce susceptibility to infection by SARS-CoV-2 and potentially disease severity. *The Lancet Infectious Diseases* 21(9): e296–e301.
- Staden R (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research* 6(7): 2601–2610.
- Stadler T, Kouyos R, Wyl V von, Yerly S, Böni J, Bürgisser P, Klimkait T, Joos B, Rieder P, Xie D, et al. (2012). Estimating the Basic Reproductive Number from Viral Sequence Data. *Molecular Biology and Evolution* 29(1): 347–357.

- Stadler T, Kühnert D, Rasmussen DA, and Plessis L du (2014). Insights into the Early Epidemic Spread of Ebola in Sierra Leone Provided by Viral Sequence Data. *PLoS Currents* 6: ecurrents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f.
- Steel MA and Penny D (1993). Distributions of Tree Comparison Metrics-Some New Results. *Systematic Biology* 42(2): 126.
- Steinbrück L, Klingen TR, and McHardy AC (2014). Computational Prediction of Vaccine Strains for Human Influenza A (H3N2) Viruses. *Journal of Virology* 88(20): 12123–12132.
- Storey N, Brown JR, Pereira RPA, O’Sullivan DM, Huggett JF, Williams R, Breuer J, and Harris KA (2021). Single base mutations in the nucleocapsid gene of SARS-CoV-2 affects amplification efficiency of sequence variants and may lead to assay failure. *Journal of Clinical Virology Plus* 1(3): 100037.
- Tagini F and Greub G (2017). Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review. *European Journal of Clinical Microbiology & Infectious Diseases* 36:11 36(11): 2007–2020.
- Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, and Prins P (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31(12): 2032–2034.
- Taype CA, Agapito JC, Accinelli RA, Espinoza JR, Godreuil S, Goodman SJ, Bañuls AL, and Shaw MA (2012). Genetic diversity, population structure and drug resistance of *Mycobacterium tuberculosis* in Peru. *Infection, Genetics and Evolution* 12(3): 577–585.
- The CRyPTIC Consortium and 100000 Genomes Project (2018). Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *New England Journal of Medicine* 379(15): 1403–1415.
- Therneau TM and Grambsch PM (2000). Modeling Survival Data: Extending the Cox Model. Springer.
- Thiry E, Meurens F, Muylkens B, McVoy M, Gogev S, Thiry J, Vanderplasschen A, Epstein A, Keil G, and Schynts F (2005). Recombination in alphaherpesviruses. *Reviews in Medical Virology* 15(2): 89–103.

- Thomer AK, Twidale MB, Guo J, and Yoder MJ (2016). Picard Tools. *Conference on Human Factors in Computing Systems - Proceedings*.
- Thorne JL, Kishino H, and Painter IS (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* 15(12): 1647–1657.
- Tom MR and Mina MJ (2020). To Interpret the SARS-CoV-2 Test, Consider the Cycle Threshold Value. *Clinical Infectious Diseases* 71(16): 2252–2254.
- Tong SYC, Davis JS, Eichenberger E, Holland TL, and Fowler VG (2015). Staphylococcus aureus Infections: Epidemiology, Pathophysiology, Clinical Manifestations, and Management. *Clinical Microbiology Reviews* 28(3): 603–661.
- Tonkin-Hill G, Martincorena I, Amato R, Lawson AR, Gerstung M, Johnston I, Jackson DK, Park N, Lensing SV, Quail MA, et al. (2021). Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* 10: e66857.
- Torres Ortiz A, Coronel J, Vidal JR, Bonilla C, Moore DAJ, Gilman RH, Balloux F, Kon OM, Didelot X, and Grandjean L (2021). Genomic signatures of pre-resistance in Mycobacterium tuberculosis. *Nature communications* 12(1): 7312.
- Torres Ortiz A, Kendall M, Storey N, Hatcher J, Dunn H, Roy S, Williams R, Williams C, Goldstein RA, Didelot X, et al. (2022). Within-host diversity improves phylogenetic and transmission reconstruction of SARS-CoV-2 outbreaks. *bioRxiv*: 2022.06.07.495142.
- Trunfio M, Calcagno A, Bonora S, and Di Perri G (2021). Lowering SARS-CoV-2 viral load might affect transmission but not disease severity in secondary cases. *The Lancet Infectious Diseases* 21(7): 914–915.
- Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Salmoniere YOLG de la, Aman K, Kato-Maeda M, and Small PM (2004). Functional and evolutionary genomics of Mycobacterium tuberculosis: Insights from genomic deletions in 100 strains. *Proceedings of the National Academy of Sciences* 101(14): 4865–4870.
- Tufariello JM, Chapman JR, Kerantzas CA, Wong KW, Vilchèze C, Jones CM, Cole LE, Tinaztepe E, Thompson V, Fenyö D, et al. (2016). Separable roles for Mycobac-

- terium tuberculosis ESX-3 effectors in iron acquisition and virulence. *Proceedings of the National Academy of Sciences of the United States of America* 113(3): E348–57.
- Turett GS, Fazal BA, Justman JE, Alland D, Duncalf RM, and Telzak EE (1997). Exogenous Reinfection with Multidrug-Resistant *Mycobacterium tuberculosis*. *Clinical Infectious Diseases* 24(3): 513–514.
- Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A, Hall G, Choi JH, Lapointe H, Kamelian K, Smith AD, et al. (2020). Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv*.
- Ugarte-Gil C, Alisjahbana B, Ronacher K, Riza AL, Koesoemadinata RC, Malherbe ST, Cioboata R, Llontop JC, Kleynhans L, Lopez S, et al. (2020). Diabetes mellitus among pulmonary tuberculosis patients from 4 tuberculosis-endemic countries: The tandem study. *Clinical Infectious Diseases* 70(5): 780–788.
- Valesano AL, Rumfelt KE, Dimcheff DE, Blair CN, Fitzsimmons WJ, Petrie JG, Martin ET, and Luring AS (2021). Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *PLOS Pathogens* 17(4): e1009499.
- Varble A, Albrecht RA, Backes S, Crumiller M, Bouvier NM, Sachs D, García-Sastre A, and Tenover BR (2014). Influenza A virus transmission bottlenecks are defined by infection route and recipient host. *Cell Host and Microbe* 16(5): 691–700.
- Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, Anyaneji UJ, Bester PA, Boni MF, Chand M, et al. (2022). Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* 2022: 1–10.
- Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, Southgate J, Johnson R, Jackson B, Nascimento FF, et al. (2021a). Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* 184(1): 64–75.e11.
- Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, Hinsley WR, Laydon DJ, Dabrera G, O'Toole Á, et al. (2021b). Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* 593(7858): 266–269.

- Volz EM, Wiuf C, Grad YH, Frost SDW, Dennis AM, and Didelot X (2020). Identification of hidden population structure in time-scaled phylogenies. *Systematic Biology*.
- Volz EM and Didelot X (2018). Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. *Systematic Biology*.
- Volz EM, Koelle K, and Bedford T (2013). Viral Phylodynamics. *PLOS Computational Biology* 9(3): e1002947.
- Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ, et al. (2015). Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *The Lancet Infectious Diseases* 15(10): 1193–1202.
- Walter KS, Colijn C, Cohen T, Mathema B, Liu Q, Bowers J, Engelthaler DM, Narechania A, Lemmer D, Croda J, et al. (2020). Genomic variant-identification methods may alter mycobacterium tuberculosis transmission inferences. *Microbial Genomics* 6(8): 1–16.
- Wang L, Didelot X, Yang J, Wong G, Shi Y, Liu W, Gao GF, and Bi Y (2020). Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. *Nature communications* 11(1): 5006.
- Warren JL, Grandjean L, Moore DAJ, Lithgow A, Coronel J, Sheen P, Zelner JL, Andrews JR, and Cohen T (2018). Investigating spillover of multidrug-resistant tuberculosis from a prison: a spatial and molecular epidemiological analysis. *BMC Medicine* 16(1): 122.
- Welch JJ and Bromham L (2005). Molecular dating when rates vary. *Trends in Ecology & Evolution* 20(6): 320–327.
- Werngren J and Hoffner SE (2003). Drug-susceptible Mycobacterium tuberculosis Beijing genotype does not develop mutation-conferred resistance to rifampin at an elevated rate. *Journal of Clinical Microbiology* 41(4): 1520–1524.

- Worby CJ, Lipsitch M, and Hanage WP (2014). Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. *PLoS Computational Biology* 10(3): e1003549.
- World Health Organization (2009). Global tuberculosis report 2009.
- (2020). Global tuberculosis report 2020. World Health Organization.
- (2021). Global tuberculosis report 2021. World Health Organization.
- Wright GD (2007). The antibiotic resistome: the nexus of chemical and genetic diversity. *Nature Reviews Microbiology* 5(3): 175–186.
- Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579(7798): 265–269.
- Wymant C, Blanquart F, Golubchik T, Gall A, Bakker M, Bezemer D, Croucher NJ, Hall M, Hillebrecht M, Ong SH, et al. (2018a). Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus Evolution* 4(1): 7.
- Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, Cesare M de, Gall A, Cornelissen M, Fraser C, and STOP-HCV Consortium {and} The BEEHIVE Collaboration TMPC (2018b). PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. *Molecular Biology and Evolution* 35(3): 719–733.
- Xin H, Li Y, Wu P, Li Z, Lau EHY, Qin Y, Wang L, Cowling BJ, Tsang TK, and Li Z (2022). Estimating the Latent Period of Coronavirus Disease 2019 (COVID-19). *Clinical Infectious Diseases* 74(9): 1678–1681.
- Xue KS, Moncla LH, Bedford T, and Bloom JD (2018). Within-Host Evolution of Human Influenza Virus. *Trends in Microbiology* 26(9): 781–793.
- Yang C, Luo T, Shen X, Wu J, Gan M, Xu P, Wu Z, Lin S, Tian J, Liu Q, et al. (2017). Transmission of multidrug-resistant *Mycobacterium tuberculosis* in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *The Lancet. Infectious diseases* 17(3): 275–284.

- Yang Y, Lin Y, and Qiao L (2018). Direct MALDI-TOF MS Identification of Bacterial Mixtures. *Analytical Chemistry* 90(17): 10400–10408.
- Yang Z, Kumar S, and Nei M (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141(4): 1641–50.
- Yoder AD and Yang Z (2000). Estimation of Primate Speciation Dates Using Local Molecular Clocks. *Molecular Biology and Evolution* 17(7): 1081–1090.
- Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, and Ballegooijen WM van (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B: Biological Sciences* 279(1728): 444–450.
- Ypma RJ, Ballegooijen WM van, and Wallinga J (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* 195(3): 1055–1062.
- Zetola NM, Shin SS, Tumedi KA, Moeti K, Ncube R, Nicol M, Collman RG, Klausner JD, and Modongoa C (2014). Mixed Mycobacterium tuberculosis complex infections and false-negative results for rifampin resistance by genexpert MTB/RIF are associated with poor clinical outcomes. *Journal of Clinical Microbiology* 52(7): 2422–2429.
- Zheng C, Li S, Luo Z, Pi R, Sun H, He Q, Tang K, Luo M, Li Y, Couvin D, et al. (2015). Mixed Infections and Rifampin Heteroresistance among Mycobacterium tuberculosis Clinical Isolates. *Journal of Clinical Microbiology* 53(7): 2138–2147.
- Zong Z, Huo F, Shi J, Jing W, Ma Y, Liang Q, Jiang G, Dai G, Huang H, and Pang Y (2018). Relapse versus reinfection of recurrent tuberculosis patients in a national tuberculosis specialized hospital in Beijing, China. *Frontiers in Microbiology* 9 (AUG): 1858.
- Zwart MP and Elena SF (2015). Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. *Annual Review of Virology* 2: 161–179.

Appendix A

Appendix

Chapter 2

Table A2.1: Assembly metadata.

Assembly name	BioSample	BioProject	Date	Sequencing technology	Lineage
ASM858v1	SAMN02603992	PRJNA223	8/4/04	sanger	Lineage 2.2
ASM944v1	SAMEA2272684	PRJEA18059	1/8/07	sanger	M. bovis
ASM1068v1	SAMD00060950	PRJDA31211	3/9/09	sanger	M. bovis
ASM1614v1	SAMN02603128	PRJNA18883	5/31/07	NA	Lineage 4.9
ASM1692v1	SAMN00103286	PRJNA15642	6/7/07	NA	Lineage 4.3
ASM2362v1	SAMN00763683	PRJNA21055	7/9/09	NA	Lineage 4.3
ASM15368v2	SAMN03081423	PRJNA17353	9/11/13	sanger	Lineage 4.1
ASM15458v2	SAMN03081422	PRJNA21053	4/7/11	NA	Lineage 4.3
ASM15460v2	SAMN00103288	PRJNA21057	6/19/12	illumina	Lineage 4.3
ASM19318v2	SAMN02595346	PRJNA39283	8/19/15	sanger;illumina;pacbio	Lineage 2.2
ASM19407v3	SAMN00771609	PRJNA63839	3/14/14	illumina	M. bovis
ASM19583v2	SAMEA20450668	PRJNA89	1/19/17	sanger	M. bovis
ASM19595v2	SAMEA3138326	PRJNA224	2/1/13	sanger	Lineage 4.9
ASM22443v1	SAMN02603755	PRJNA43171	8/25/11	sanger;roche.454	Lineage 4.3
ASM23472v1	SAMN02604104	PRJNA45811	11/9/11	sanger;roche.454	M. bovis
ASM25335v1	SAMEA2272165	PRJEA68095	7/5/11	sanger;illumina	Lineage 6.1
ASM27034v1	SAMN02603010	PRJNA19585	7/5/11	sanger;illumina	Lineage 2.2
ASM27036v1	SAMN02603011	PRJNA19583	7/5/11	sanger;illumina	Lineage 2.2
ASM27708v1	SAMN02603736	PRJNA73717	3/20/12	roche.454	Lineage 4.3
ASM27710v1	SAMN02603737	PRJNA73719	3/20/12	roche.454	Lineage 1.2
ASM27773v2	SAMN03081436	PRJNA37301	11/26/13	sanger;roche.454;illumina	Lineage 4.9

Chapter A. Appendix

ASM33144v1	SAMEA2272730	PRJEA82273	1/8/13	sanger;roche.454	Lineage 4.1
ASM33871v2	SAMN02603488	PRJNA170028	3/8/13	roche.454;illumina	M. bovis
ASM35020v1	SAMD00061050	PRJDB66	3/1/12	sanger;roche.454;illumina	Lineage 4.1
ASM38990v1	SAMN02603650	PRJNA192916	5/8/13	illumina	Lineage 4.9
ASM38992v1	SAMN02603652	PRJNA193386	5/8/13	illumina	Lineage 4.5
ASM38994v1	SAMN02603653	PRJNA193387	5/8/13	illumina	Lineage 4.9
ASM40061v1	SAMN02603646	PRJNA66069	5/20/13	roche.454;illumina	Lineage 2.2
ASM42212v1	SAMN02603180	PRJNA210717	7/11/13	illumina	Lineage 1.1
ASM57212v1	SAMN03081429	PRJNA66061	2/18/14	NA	Lineage 2.2
ASM57215v1	SAMN03081430	PRJNA66063	2/18/14	NA	Lineage 2.2
ASM57217v1	SAMN03081431	PRJNA66065	2/18/14	NA	Lineage 2.2
ASM57219v1	SAMN03081432	PRJNA66071	2/18/14	NA	Lineage 2.2
ASM69847v1	SAMN02803802	PRJNA178919	6/5/14	sanger;illumina	Lineage 2.2
ASM70666v1	SAMN02767600	PRJNA246436	6/13/14	roche.454;illumina	Lineage 2.2
ASM73844v1	SAMN02943514	PRJNA256478	8/7/14	ion.torrent	Lineage 4.4
ASM73847v1	SAMN02943515	PRJNA256478	8/7/14	ion.torrent	Lineage 4.4
ASM75652v1	SAMN02887314	PRJNA253748	9/22/14	roche.454	Lineage 2.2
ASM75654v1	SAMN02874015	PRJNA253748	9/22/14	sanger;roche.454;illumina	Lineage 1.2
MT49-02	SAMEA3865285	PRJEB4916	11/3/14	NA	Lineage 2.2
ASM82708v1	SAMN03287041	PRJNA232612	1/21/15	illumina	Lineage 4.9
ASM82899v1	SAMD00000334	PRJDB1582	12/23/14	pacbio	Lineage 4.9
ASM93432v3	SAMN03288261	PRJNA273066	8/5/16	illumina	M. bovis
ASM95415v1	SAMN03258416	PRJNA269967	3/17/15	pacbio	Lineage 2.2
ASM96728v1	SAMEA4532327	PRJNA18279	7/25/11	sanger	Lineage 4.9
ASM104325v1	SAMN02864173	PRJNA251957	6/30/15	sanger;roche.454;illumina	M. bovis
ASM107861v1	SAMN03761171	PRJNA285833	7/17/15	roche.454;illumina;pacbio	M. bovis
ASM127455v1	SAMN02940916	PRJNA256163	8/26/15	sanger;roche.454	M. bovis
ASM127556v2	SAMN03921880	PRJNA291131	8/12/16	ion.torrent;illumina	Lineage 2.2
ASM148390v1	SAMN04364424	PRJNA306822	1/4/16	sanger;roche.454;IonTorrent	M. bovis
ASM154467v1	SAMN03253054	PRJNA270004	2/2/16	pacbio	Lineage 4.9
ASM154470v1	SAMN03253058	PRJNA270004	2/2/16	pacbio	Lineage 4.9
ASM154473v1	SAMN03253093	PRJNA270004	2/2/16	pacbio	M. bovis
ASM154477v1	SAMN03257089	PRJNA270004	2/2/16	pacbio	M. bovis
ASM154481v1	SAMN03257090	PRJNA270004	2/2/16	pacbio	M. microti
ASM154485v1	SAMN03257091	PRJNA270004	2/2/16	pacbio	Lineage 6.1
ASM154489v1	SAMN03257092	PRJNA270004	2/2/16	pacbio	Lineage 2.2
ASM154493v1	SAMN03257093	PRJNA270004	2/2/16	pacbio	Lineage 2.2
ASM154495v1	SAMN03257094	PRJNA270004	2/2/16	pacbio	Lineage 4.5
ASM154498v1	SAMN03257095	PRJNA270004	2/2/16	pacbio	Lineage 4.4
ASM154501v1	SAMN03257096	PRJNA270004	2/2/16	pacbio	Lineage 4.2
ASM154505v1	SAMN03257097	PRJNA270004	2/2/16	pacbio	Lineage 3
ASM158038v1	SAMN04517448	PRJNA313242	3/7/16	illumina	M. bovis

ASM159764v2	SAMN04492771	PRJNA312026	12/11/19	sanger;roche.454;illumina	Lineage 4.9
ASM170243v1	SAMN05447655	PRJNA335538	8/10/16	illumina	Lineage 2.2
ASM170826v1	SAMN05447658	PRJNA335537	8/19/16	illumina	Lineage 2.2
ASM175086v1	SAMN03301665	PRJNA273899	10/3/16	sanger;illumina	Lineage 2.2
ASM185525v1	SAMN04293604	PRJNA304069	10/27/16	sanger;roche.454	Lineage 2.2
ASM187014v1	SAMN05933692	PRJNA349996	11/7/16	pacbio	Lineage 2.2
ASM189576v1	SAMN06010135	PRJNA343736	12/9/16	illumina	Lineage 4.3
ASM189578v1	SAMN06010493	PRJNA343736	12/9/16	illumina	Lineage 4.3
ASM189580v1	SAMN06010122	PRJNA343736	12/9/16	illumina	Lineage 4.3
ASM189582v1	SAMN06010171	PRJNA343736	12/9/16	illumina	Lineage 4.1
ASM189584v1	SAMN06010490	PRJNA343736	12/9/16	illumina	Lineage 4.3
ASM189586v1	SAMN06010343	PRJNA343736	12/9/16	illumina	Lineage 4.3
ASM192248v1	SAMN06141171	PRJNA357378	12/27/16	illumina	Lineage 4.5
ASM193872v1	SAMN05414673	PRJNA329548	1/6/17	pacbio	Lineage 4.9
ASM207277v2	SAMN06622696	PRJNA379898	4/6/17	pacbio	Lineage 4.5
ASM211675v1	SAMN05833360	PRJNA269967	5/1/17	illumina;pacbio	Lineage 2.2
ASM211677v1	SAMN05833361	PRJNA269967	5/1/17	illumina;pacbio	Lineage 2.2
ASM211679v1	SAMN05833362	PRJNA269967	5/1/17	illumina;pacbio	Lineage 2.2
ASM211681v1	SAMN05833364	PRJNA269967	5/1/17	illumina;pacbio	Lineage 2.2
ASM211683v1	SAMN05833363	PRJNA269967	5/1/17	illumina;pacbio	Lineage 2.2
ASM211685v1	SAMN05833359	PRJNA269967	5/1/17	illumina;pacbio	Lineage 2.2
ASM220823v1	SAMN06622697	PRJNA379898	6/26/17	pacbio	Lineage 2.2
ASM235601v1	SAMD00065354	PRJDB5220	12/1/16	pacbio	Lineage 4.3
ASM235625v1	SAMD00076259	PRJDB5608	4/25/17	sanger;illumina;pacbio	Lineage 1.1
ASM235793v1	SAMD00076256	PRJDB5607	5/10/17	sanger;illumina;pacbio	Lineage 2.2
ASM235795v1	SAMD00076257	PRJDB5607	5/10/17	sanger;illumina;pacbio	Lineage 2.2
ASM235797v1	SAMD00076258	PRJDB5607	5/10/17	sanger;illumina;pacbio	Lineage 2.2
ASM244687v1	SAMN07664823	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244689v1	SAMN07664824	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244691v1	SAMN07664825	PRJNA407834	10/5/17	illumina	Lineage 4
ASM244693v1	SAMN07664826	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244695v1	SAMN07664827	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244697v1	SAMN07664828	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244699v1	SAMN07664829	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244701v1	SAMN07664830	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244703v1	SAMN07664831	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244705v1	SAMN07664832	PRJNA407834	10/5/17	illumina	Lineage 4
ASM244707v1	SAMN07664833	PRJNA407834	10/5/17	illumina	Lineage 4
ASM244709v1	SAMN07664834	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244711v1	SAMN07664835	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244713v1	SAMN07664836	PRJNA407834	10/5/17	illumina	Lineage 2.2
ASM244715v1	SAMN07664837	PRJNA407834	10/5/17	illumina	Lineage 4.3

ASM244717v1	SAMN07664838	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244719v1	SAMN07664839	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244721v1	SAMN07664840	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244723v1	SAMN07664841	PRJNA407834	10/5/17	illumina	Lineage 4
ASM244725v1	SAMN07664842	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244727v1	SAMN07664843	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244729v1	SAMN07664844	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244731v1	SAMN07664845	PRJNA407834	10/5/17	illumina	Lineage 2.2
ASM244733v1	SAMN07664846	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244735v1	SAMN07664847	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244737v1	SAMN07664848	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244739v1	SAMN07664849	PRJNA407834	10/5/17	illumina	Lineage 2.2
ASM244741v1	SAMN07664850	PRJNA407834	10/5/17	illumina	Lineage 4.7
ASM244743v1	SAMN07664851	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244745v1	SAMN07664852	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244747v1	SAMN07664853	PRJNA407834	10/5/17	illumina	Lineage 4
ASM244749v1	SAMN07664854	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244751v1	SAMN07664855	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244753v1	SAMN07664856	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244755v1	SAMN07664857	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244757v1	SAMN07664858	PRJNA407834	10/5/17	illumina	Lineage 4.8
ASM244759v1	SAMN07664859	PRJNA407834	10/5/17	illumina	Lineage 2.2
ASM244761v1	SAMN07664860	PRJNA407834	10/5/17	illumina	Lineage 4
ASM244763v1	SAMN07664861	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244765v1	SAMN07664862	PRJNA407834	10/5/17	illumina	Lineage 4
ASM244767v1	SAMN07664863	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244769v1	SAMN07664864	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244771v1	SAMN07664865	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244773v1	SAMN07664866	PRJNA407834	10/5/17	illumina	Lineage 4.8
ASM244775v1	SAMN07664867	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244777v1	SAMN07664868	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244779v1	SAMN07664869	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244781v1	SAMN07664870	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244783v1	SAMN07664871	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244785v1	SAMN07664872	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244787v1	SAMN07664873	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244789v1	SAMN07664874	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244791v1	SAMN07664875	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244793v1	SAMN07664876	PRJNA407834	10/5/17	illumina	Lineage 2.2
ASM244795v1	SAMN07664877	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244797v1	SAMN07664878	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244799v1	SAMN07664879	PRJNA407834	10/5/17	illumina	Lineage 4.3

ASM244801v1	SAMN07664880	PRJNA407834	10/5/17	illumina	Lineage 4.7
ASM244803v1	SAMN07664881	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244805v1	SAMN07664882	PRJNA407834	10/5/17	illumina	Lineage 4
ASM244807v1	SAMN07664883	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244809v1	SAMN07664884	PRJNA407834	10/5/17	illumina	Lineage 4.8
ASM244811v1	SAMN07664885	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244813v1	SAMN07664886	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244815v1	SAMN07664887	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM244817v1	SAMN07664888	PRJNA407834	10/5/17	illumina	Lineage 4.3
ASM244819v1	SAMN07664889	PRJNA407834	10/5/17	illumina	Lineage 2.2
ASM244821v1	SAMN07664890	PRJNA407834	10/5/17	illumina	Lineage 4.1
ASM288614v1	SAMN08199254	PRJNA422870	1/17/18	illumina	Lineage 4.1
ASM288616v1	SAMN08199260	PRJNA422870	1/17/18	illumina	Lineage 4.8
ASM288619v1	SAMN08199302	PRJNA422870	1/17/18	illumina	Lineage 4.3
ASM288622v1	SAMN08199276	PRJNA422870	1/17/18	illumina	Lineage 4.3
ASM288633v1	SAMN08199305	PRJNA422870	1/17/18	illumina	Lineage 4.3
ASM288640v1	SAMN08199301	PRJNA422870	1/17/18	illumina	Lineage 4.1
ASM288650v1	SAMN08199277	PRJNA422870	1/17/18	illumina	Lineage 4.3
ASM288658v1	SAMN08199266	PRJNA422870	1/17/18	illumina	Lineage 4.3
ASM288668v1	SAMN08199347	PRJNA422870	1/17/18	illumina	Lineage 4.3
ASM288677v1	SAMN08199271	PRJNA422870	1/17/18	illumina	Lineage 4.3
ASM288686v1	SAMN08199268	PRJNA422870	1/17/18	illumina	Lineage 4.1
ASM288694v1	SAMN08199265	PRJNA422870	1/17/18	illumina	Lineage 4.1
ASM288706v1	SAMN08199267	PRJNA422870	1/17/18	illumina	Lineage 4.8
ASM288714v1	SAMN08199303	PRJNA422870	1/17/18	illumina	Lineage 4.8
ASM288725v1	SAMN08199306	PRJNA422870	1/17/18	illumina	Lineage 4.3
ASM288733v1	SAMN08199348	PRJNA422870	1/17/18	illumina	Lineage 4.3
ASM289518v2	SAMN07125453	PRJNA386696	3/19/18	illumina;nanopore	Lineage 2.2
ASM297547v1	SAMN08321034	PRJNA428763	3/7/18	illumina;pacbio	M. bovis
ASM300611v1	SAMN07414820	PRJNA395787	3/16/18	pacbio	Lineage 2.2
ASM300613v1	SAMN07414819	PRJNA395786	3/16/18	pacbio	Lineage 2.2
ASM326500v1	SAMN09453176	PRJNA476697	6/26/18	illumina	Lineage 2.2
ASM328712v1	SAMN06292592	PRJNA369711	7/3/18	pacbio	Lineage 2.2
ASM328714v1	SAMN06292590	PRJNA369711	7/3/18	pacbio	Lineage 4.1
ASM328716v1	SAMN06292065	PRJNA369711	7/3/18	pacbio	Lineage 4.5
ASM328718v1	SAMN06292067	PRJNA369711	7/3/18	pacbio	Lineage 2.2
ASM329365v1	SAMN08973501	PRJNA453104	7/9/18	illumina	Lineage 1.2
ASM515578v1	SAMN10219709	PRJNA494982	5/6/19	illumina;pacbio	M. bovis
ASM515610v1	SAMN10219708	PRJNA494982	5/6/19	illumina;pacbio	M. bovis
ASM783393v1	SAMN12126281	PRJNA550393	8/5/19	pacbio	Lineage 4.1
ASM876167v1	SAMN12816907	PRJNA573545	10/1/19	illumina;pacbio	Lineage 4.4
ASM966487v1	SAMN13191639	PRJNA556438	11/17/19	illumina;pacbio	Lineage 1.1

Chapter A. Appendix

ASM973021v1	SAMN11056466	PRJNA231221	12/5/19	illumina;pacbio	Lineage 1.2
ASM973023v1	SAMN11056465	PRJNA231221	12/5/19	illumina;pacbio	Lineage 3
ASM976267v2	SAMN10417149	PRJNA505382	2/24/20	illumina;pacbio	Lineage 2.2
ASM1104445v1	SAMN14150054	PRJNA607885	3/1/20	illumina;nanopore	Lineage 2.2
ASM1292376v1	SAMN13736268	PRJNA598991	4/28/20	illumina;pacbio	Lineage 8
ASM1301038v1	SAMN14598146	PRJNA625312	5/11/20	illumina;nanopore	Lineage 4.3
ASM1326763v1	SAMN11056472	PRJNA231221	6/4/20	illumina;pacbio	Lineage 4.9
ASM1326765v1	SAMN11056471	PRJNA231221	6/4/20	illumina;pacbio	Lineage 2.2
ASM1488462v1	SAMN12325246	PRJNA555636	10/13/20	pacbio	Lineage 4.3
ASM1488464v1	SAMN12325211	PRJNA555636	10/13/20	pacbio	Lineage 2.2
ASM1488466v1	SAMN12325208	PRJNA555636	10/13/20	pacbio	Lineage 2.2
ASM1489920v1	SAMN12325280	PRJNA555636	10/20/20	pacbio	Lineage 4.1
ASM1489922v1	SAMN12325279	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1489924v1	SAMN12325278	PRJNA555636	10/20/20	pacbio	Lineage 1.1
ASM1489926v1	SAMN12325276	PRJNA555636	10/20/20	pacbio	Lineage 1.1
ASM1489928v1	SAMN12325277	PRJNA555636	10/20/20	pacbio	Lineage 1.1
ASM1489930v1	SAMN12325275	PRJNA555636	10/20/20	pacbio	Lineage 1.1
ASM1489932v1	SAMN12325274	PRJNA555636	10/20/20	pacbio	Lineage 1.1
ASM1489934v1	SAMN12325273	PRJNA555636	10/20/20	pacbio	Lineage 1.1
ASM1489936v1	SAMN12325272	PRJNA555636	10/20/20	pacbio	Lineage 1.2
ASM1489938v1	SAMN12325271	PRJNA555636	10/20/20	pacbio	Lineage 2.1
ASM1489940v1	SAMN12325270	PRJNA555636	10/20/20	pacbio	Lineage 1.1
ASM1489942v1	SAMN12325269	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1489944v1	SAMN12325268	PRJNA555636	10/20/20	pacbio	Lineage 1.2
ASM1489946v1	SAMN12325267	PRJNA555636	10/20/20	pacbio	Lineage 1.2
ASM1489948v1	SAMN12325266	PRJNA555636	10/20/20	pacbio	Lineage 1.1
ASM1489950v1	SAMN12325265	PRJNA555636	10/20/20	pacbio	Lineage 7
ASM1489952v1	SAMN12325264	PRJNA555636	10/20/20	pacbio	Lineage 4.8
ASM1489954v1	SAMN12325263	PRJNA555636	10/20/20	pacbio	Lineage 4.1
ASM1489956v1	SAMN12325262	PRJNA555636	10/20/20	pacbio	Lineage 4.4
ASM1489958v1	SAMN12325261	PRJNA555636	10/20/20	pacbio	Lineage 4.8
ASM1489960v1	SAMN12325260	PRJNA555636	10/20/20	pacbio	Lineage 4.3
ASM1489962v1	SAMN12325259	PRJNA555636	10/20/20	pacbio	Lineage 4.7
ASM1489964v1	SAMN12325258	PRJNA555636	10/20/20	pacbio	Lineage 4.3
ASM1489966v1	SAMN12325257	PRJNA555636	10/20/20	pacbio	Lineage 1.2
ASM1489968v1	SAMN12325256	PRJNA555636	10/20/20	pacbio	Lineage 4.3
ASM1489970v1	SAMN12325255	PRJNA555636	10/20/20	pacbio	Lineage 3.1
ASM1489972v1	SAMN12325254	PRJNA555636	10/20/20	pacbio	Lineage 4.1
ASM1489974v1	SAMN12325253	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1489976v1	SAMN12325252	PRJNA555636	10/20/20	pacbio	Lineage 4.1
ASM1489978v1	SAMN12325251	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1489980v1	SAMN12325250	PRJNA555636	10/20/20	pacbio	Lineage 4.1

ASM1489982v1	SAMN12325249	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1489984v1	SAMN12325248	PRJNA555636	10/20/20	pacbio	Lineage 4.3
ASM1489986v1	SAMN12325247	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1489988v1	SAMN12325245	PRJNA555636	10/20/20	pacbio	Lineage 4.3
ASM1489990v1	SAMN12325244	PRJNA555636	10/20/20	pacbio	Lineage 4.3
ASM1489992v1	SAMN12325243	PRJNA555636	10/20/20	pacbio	Lineage 4.1
ASM1489994v1	SAMN12325242	PRJNA555636	10/20/20	pacbio	Lineage 4.3
ASM1489996v1	SAMN12325241	PRJNA555636	10/20/20	pacbio	Lineage 1.2
ASM1489998v1	SAMN12325240	PRJNA555636	10/20/20	pacbio	Lineage 1.2
ASM1490000v1	SAMN12325239	PRJNA555636	10/20/20	pacbio	Lineage 1.2
ASM1490005v1	SAMN12325238	PRJNA555636	10/20/20	pacbio	Lineage 4.3
ASM1490007v1	SAMN12325237	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490009v1	SAMN12325236	PRJNA555636	10/20/20	pacbio	Lineage 4.2
ASM1490011v1	SAMN12325235	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490013v1	SAMN12325234	PRJNA555636	10/20/20	pacbio	Lineage 4.2
ASM1490015v1	SAMN12325233	PRJNA555636	10/20/20	pacbio	Lineage 4.1
ASM1490017v1	SAMN12325232	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490019v1	SAMN12325231	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490021v1	SAMN12325230	PRJNA555636	10/20/20	pacbio	Lineage 4.2
ASM1490023v1	SAMN12325229	PRJNA555636	10/20/20	pacbio	Lineage 4.2
ASM1490025v1	SAMN12325228	PRJNA555636	10/20/20	pacbio	Lineage 4.2
ASM1490027v1	SAMN12325227	PRJNA555636	10/20/20	pacbio	Lineage 4.2
ASM1490029v1	SAMN12325226	PRJNA555636	10/20/20	pacbio	Lineage 4.2
ASM1490031v1	SAMN12325225	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490033v1	SAMN12325224	PRJNA555636	10/20/20	pacbio	Lineage 4.3
ASM1490035v1	SAMN12325223	PRJNA555636	10/20/20	pacbio	Lineage 3
ASM1490037v1	SAMN12325222	PRJNA555636	10/20/20	pacbio	Lineage 4.8
ASM1490039v1	SAMN12325221	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490041v1	SAMN12325220	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490043v1	SAMN12325219	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490045v1	SAMN12325218	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490047v1	SAMN12325217	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490049v1	SAMN12325216	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490051v1	SAMN12325215	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490055v1	SAMN12325214	PRJNA555636	10/20/20	pacbio	Lineage 3.1
ASM1490059v1	SAMN12325213	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490061v1	SAMN12325212	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490063v1	SAMN12325210	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490065v1	SAMN12325209	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490067v1	SAMN12325207	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490069v1	SAMN12325206	PRJNA555636	10/20/20	pacbio	Lineage 4
ASM1490071v1	SAMN12325205	PRJNA555636	10/20/20	pacbio	Lineage 2.2

ASM1490073v1	SAMN12325204	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490075v1	SAMN12325203	PRJNA555636	10/20/20	pacbio	Lineage 3
ASM1490077v1	SAMN12325202	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490079v1	SAMN12325201	PRJNA555636	10/20/20	pacbio	Lineage 4.8
ASM1490081v1	SAMN12325200	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490083v1	SAMN12325199	PRJNA555636	10/20/20	pacbio	Lineage 4.2
ASM1490085v1	SAMN12325198	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490089v1	SAMN12325197	PRJNA555636	10/20/20	pacbio	Lineage 1.2
ASM1490096v1	SAMN12325196	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490103v1	SAMN12325195	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490105v1	SAMN12325194	PRJNA555636	10/20/20	pacbio	Lineage 3
ASM1490107v1	SAMN12325193	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490109v1	SAMN12325192	PRJNA555636	10/20/20	pacbio	Lineage 3.1
ASM1490111v1	SAMN12325191	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1490113v1	SAMN12325190	PRJNA555636	10/20/20	pacbio	Lineage 3.1
ASM1490115v1	SAMN12325189	PRJNA555636	10/20/20	pacbio	Lineage 2.2
ASM1548280v1	SAMN16605479	PRJNA673391	11/16/20	illumina;nanopore	M. bovis
ASM1691777v1	SAMN17832565	PRJNA700485	2/22/21	illumina;nanopore	Lineage 2.2
ASM1716403v1	SAMN18056856	PRJNA704837	3/4/21	illumina	Lineage 4.9
ASM1716405v1	SAMN18056929	PRJNA704837	3/4/21	illumina	Lineage 4.9
ASM1790103v1	SAMN18576181	PRJNA718842	4/12/21	illumina	Lineage 4.9
ASM1790105v1	SAMN18576179	PRJNA718842	4/12/21	illumina	Lineage 4.9
ASM1790107v1	SAMN18576180	PRJNA718842	4/12/21	illumina	Lineage 4.9
ASM1790109v1	SAMN18576182	PRJNA718842	4/12/21	illumina	Lineage 4.9
ASM1790111v1	SAMN18576178	PRJNA718842	4/12/21	illumina	Lineage 4.9
ASM1967074v1	SAMD00322307	PRJDB11640	5/28/21	illumina;nanopore	Lineage 2.2
ASM2153515v1	SAMN23482338	PRJNA784324	1/20/22	nanopore	Lineage 4.9
ASM2155989v1	SAMN15348336	PRJNA641267	1/24/22	pacbio	Lineage 5.3
ASM2202399v1	SAMN18614372	PRJNA719670	2/7/22	illumina;pacbio	Lineage 4.4
ASM2287012v1	SAMN16406479	PRJNA668291	4/11/22	nanopore	Lineage 1.1
ASM2287014v1	SAMN16406478	PRJNA668291	4/11/22	nanopore	Lineage 7
ASM2287016v1	SAMN16406477	PRJNA668291	4/11/22	nanopore	Lineage 6.2
ASM2287018v1	SAMN16406476	PRJNA668291	4/11/22	nanopore	Lineage 6.1
ASM2287020v1	SAMN16406475	PRJNA668291	4/11/22	nanopore	Lineage 6.3
ASM2287022v1	SAMN16406474	PRJNA668291	4/11/22	nanopore	Lineage 5.1
ASM2287024v1	SAMN16406473	PRJNA668291	4/11/22	nanopore	Lineage 5.1
ASM2287026v1	SAMN16406472	PRJNA668291	4/11/22	nanopore	Lineage 5.1
ASM2287028v1	SAMN16406471	PRJNA668291	4/11/22	nanopore	Lineage 4.2
ASM2287030v1	SAMN16406470	PRJNA668291	4/11/22	nanopore	Lineage 4.3
ASM2287032v1	SAMN16406469	PRJNA668291	4/11/22	nanopore	Lineage 4.6
ASM2287034v1	SAMN16406468	PRJNA668291	4/11/22	nanopore	Lineage 3
ASM2287036v1	SAMN16406467	PRJNA668291	4/11/22	nanopore	Lineage 3.1

ASM2287038v1	SAMN16406466	PRJNA668291	4/11/22	nanopore	Lineage 3.1
ASM2287040v1	SAMN16406464	PRJNA668291	4/11/22	nanopore	Lineage 2.1
ASM2287042v1	SAMN16406463	PRJNA668291	4/11/22	nanopore	Lineage 1.1
ASM2287044v1	SAMN16406462	PRJNA668291	4/11/22	nanopore	Lineage 1.2
ASM2287108v1	SAMN27028218	PRJNA820632	4/11/22	pacbio	Lineage 2.2
ASM2287110v1	SAMN27028219	PRJNA820632	4/11/22	pacbio	Lineage 2.2
ASM2287112v1	SAMN27028220	PRJNA820632	4/11/22	pacbio	Lineage 2.2
ASM2287114v1	SAMN27028221	PRJNA820632	4/11/22	pacbio	Lineage 2.2
Mtb_DKC2.180803	SAMEA104027311	PRJEB20214	11/10/18	illumina;nanopore	Lineage 4.8
MB3601	SAMEA5803801	PRJEB33636	11/1/19	illumina;pacbio	M.bovis
MmicOV254	SAMEA7336314	PRJEB40500	1/27/21	illumina;pacbio	M. microti
MmicATCC35782	SAMEA7336315	PRJEB40500	1/27/21	illumina;pacbio	M. microti
MmicMaus3	SAMEA7336317	PRJEB40500	1/27/21	illumina;pacbio	M. microti
MmicMaus4	SAMEA7336318	PRJEB40500	1/27/21	illumina;pacbio	M. microti
Mmic942272	SAMEA7336316	PRJEB40500	1/27/21	illumina;pacbio	M. microti
mada_1-44	SAMEA13188688	PRJEB49093	3/23/22	illumina;nanopore;pacbio	Lineage 1.1
mada_125	SAMEA13188731	PRJEB49093	3/23/22	illumina;nanopore;pacbio	Lineage 2.2
mada_116	SAMEA13188685	PRJEB49093	3/23/22	illumina;nanopore;pacbio	Lineage 3.1
mada_130	SAMEA13188674	PRJEB49093	3/23/22	illumina;nanopore;pacbio	Lineage 4.1
mada_132	SAMEA13188689	PRJEB49093	3/23/22	illumina;nanopore;pacbio	Lineage 4.8
mada_104	SAMEA13188715	PRJEB49093	3/23/22	illumina;nanopore;pacbio	Lineage 1.1

Chapter 3

Table A3.1: Drug resistance profiles by project.

Drug profile ^{1,2}	Dataset				
	Population	Household	CRyPTIC	Tandem	Other
Susceptible	73.4% (1496/2037)	3.9% (7/180)	15.9% (77/483)	81.5% (327/401)	36.4% (12/33)
DR	14.1% (288/2037)	5.6% (10/180)	20.1% (97/483)	10% (40/401)	9.1% (3/33)
Rifampicin	1.6% (32/2037)	2.2% (4/180)	3.1% (15/483)	0.7% (3/401)	0% (0/33)
Isoniazid	5.2% (105/2037)	1.1% (2/180)	7.9% (38/483)	2.5% (10/401)	3% (1/33)
Ethambutol	0.5% (10/2037)	0% (0/180)	0% (0/483)	0.2% (1/401)	0% (0/33)
Streptomycin	1.3% (26/2037)	0% (0/180)	0.8% (4/483)	1.5% (6/401)	0% (0/33)
Other	5.6% (115/2037)	2.2% (4/180)	8.1% (40/483)	4.2% (20/401)	6.1% (2/33)
MDR	12.3% (251/2037)	88.3% (159/180)	59.2% (286/483)	8.5% (34/401)	51.5% (17/33)
XDR	0.1% (2/2037)	2.2% (4/180)	4.8% (23/483)	0% (0/401)	3% (1/33)

¹ Estimation of drug resistance profile using drug resistance associated SNPs **Coll2015**

² DR, drug resistance to a single antibiotic or a combination of antibiotics not included in MDR (Other); MDR, multi-drug resistance, defined as co-occurrence of isoniazid and rifampicin resistance; XDR, extensively drug resistance, defined as MDR in addition to resistance to any fluoroquinolone (ciprofloxacin, moxifloxacin, or ofloxacin) and at least one of three injectable second-line drugs (amikacin, kanamycin, or capreomycin) (World Health Organization 2020)

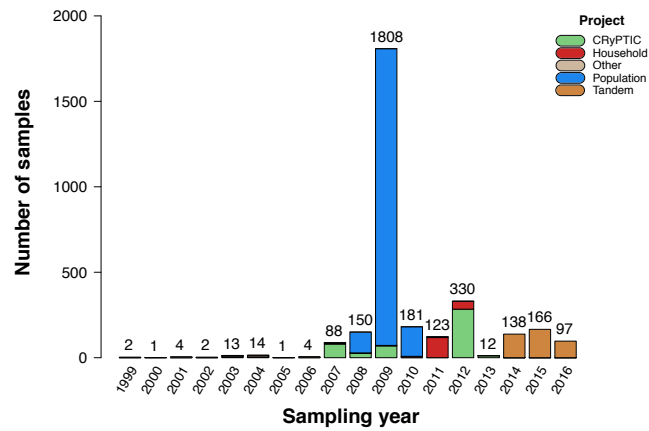


Figure A3.1: Sample cohort.
Temporal distribution of the 3134 samples included in the study.

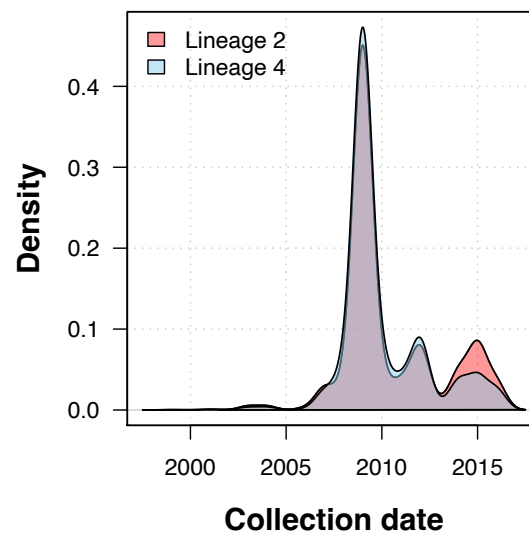


Figure A3.2: Sampling date distribution.
Distribution of sampling dates by lineage.

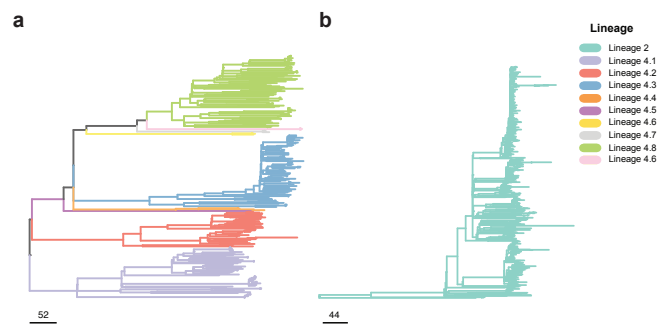


Figure A3.3: Maximum likelihood phylogeny of the Samara (Russia) data set. Phylogeny of the Samara data set for (a) lineage 4 and (b) lineage 2.

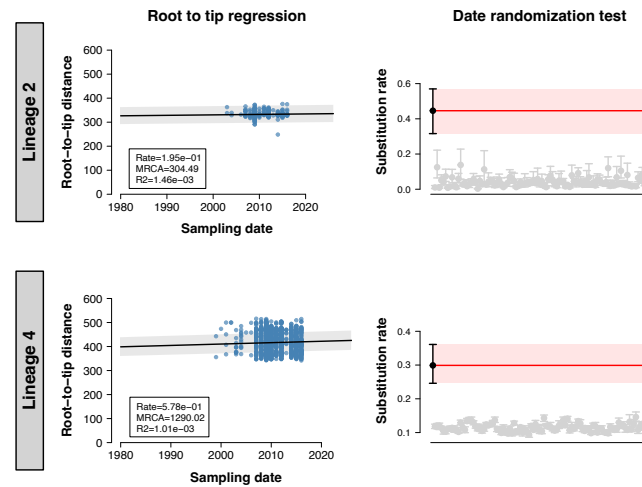


Figure A3.4: Temporal signal analysis

The presence of measurable evolution (temporal signal) within the lineage 2 (top) and lineage 4 (bottom) was tested by a root-to-tip regression (left) and by a date-randomization test (right). In the root-to-tip regression the distance from the tips to the root of the phylogenetic tree (y-axis) are plotted against the sampling dates (x-axis). The mutation rate and the time of the most recent common ancestor (MRCA) are calculated from the linear regression. The solid black line represents the best fitting regression line while the grey shaded area shows the 95% CI estimated from a Gamma distribution. The date randomization test is performed by comparing the substitution rate of the original dataset (black) with a set of randomized data sets obtained by permutating the sampling dates (grey). The points show the mean of the distribution while error bars represent the 95% CI. The mean and 95% CI of the original dataset are also highlighted by a solid red line and a red shaded area, respectively. The date randomization test is passed if the estimates from the original data do not overlap with the randomized sets.

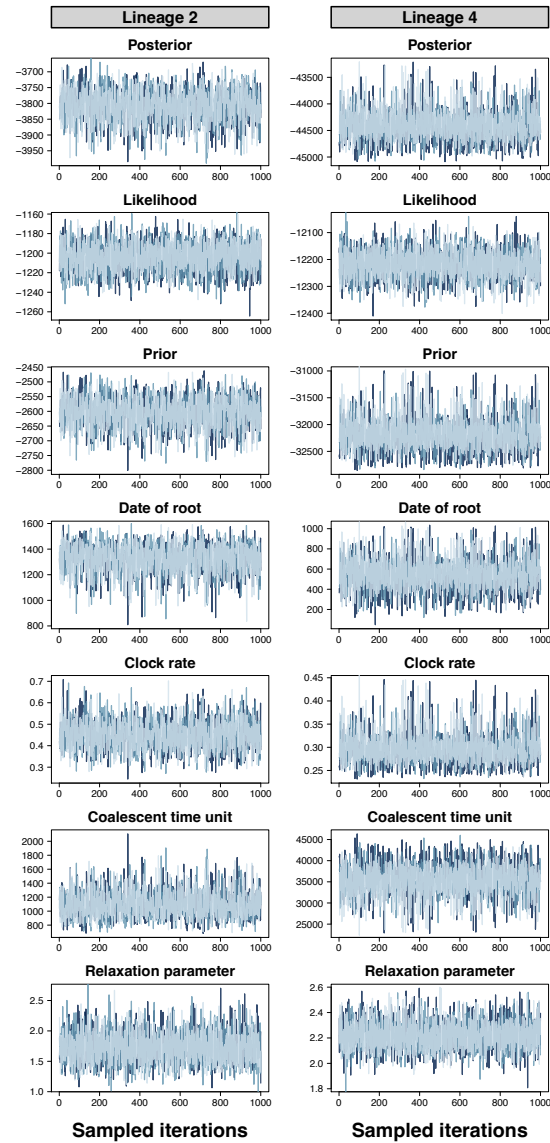


Figure A3.5: Tree dating MCMC chain convergence.

BactDating trace output for lineage 2 and lineage 4 in three independent MCMC chains. Parameters shown are posterior probabilities, likelihood, prior probabilities, date of MRCA, substitution rates, coalescent time unit and relaxation parameters. Three independent chains were run for each dataset.

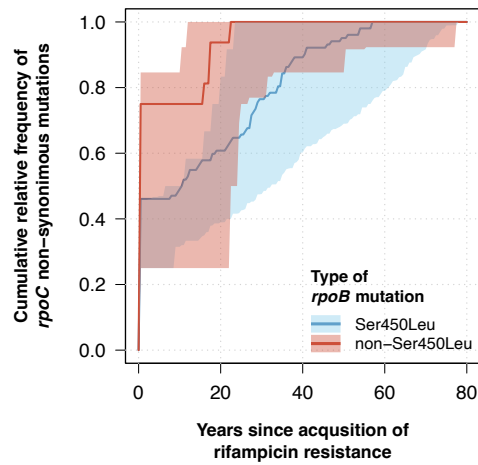


Figure A3.6: Dynamics of non-synonymous mutations in the *rpoC* gene for Ser450Leu and non-Ser450Leu *rpoB* mutations

Cumulative relative frequency of non-synonymous mutations in *rpoC* over time. The x-axis represents the years since the inferred time of rifampicin resistance (time 0). Colors show mutations for Ser450Leu (blue) and non-Ser450Leu (red) *rpoB* rifampicin resistant isolates. Dark colors represent the number of mutations for the ML tree, while the confidence interval (shaded area) is inferred by repeating the analysis in 100 bootstrap phylogenies.

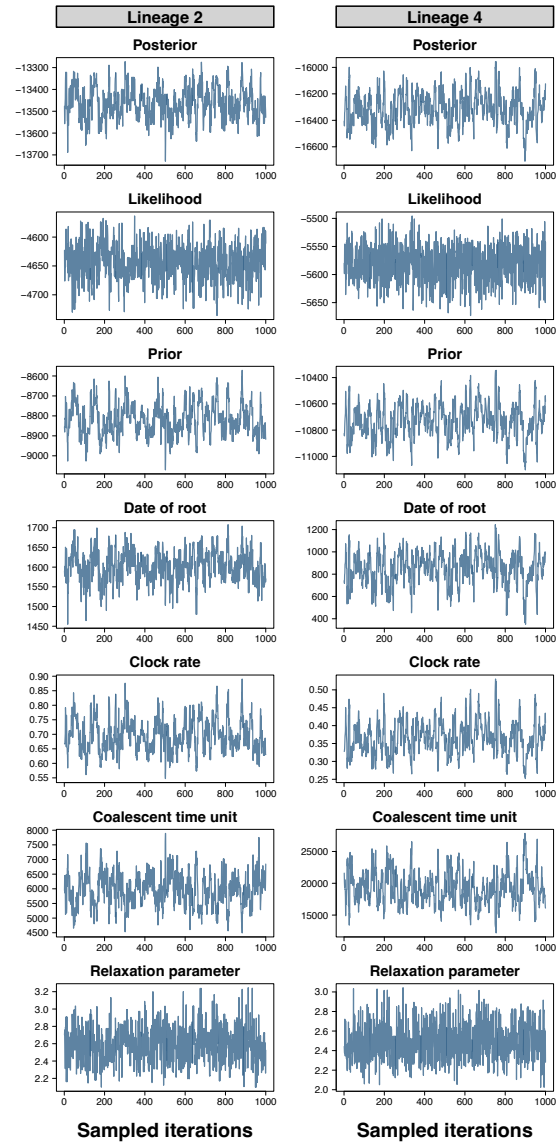


Figure A3.7: Tree dating MCMC chain convergence for the global data set. BactDating trace output for lineage 2 and lineage 4 in a global set of publicly available strains. Parameters shown are posterior probabilities, likelihood, prior probabilities, date of MRCA, substitution rates, coalescent time unit and relaxation parameters.

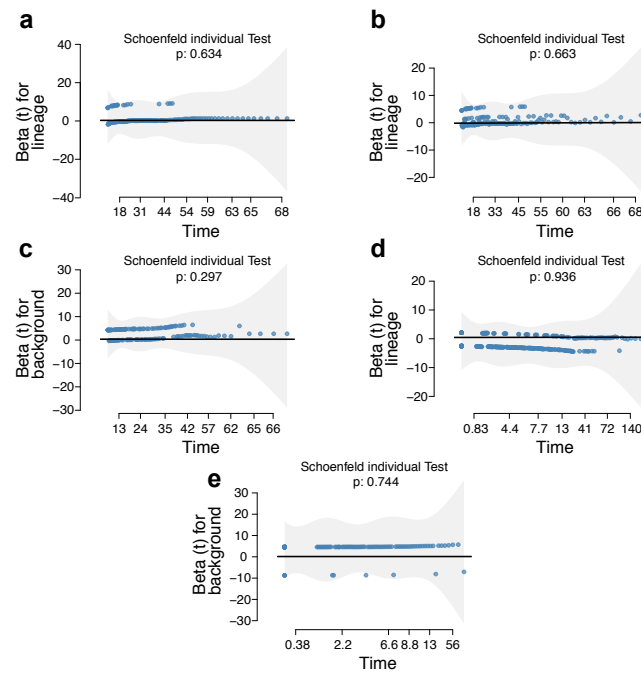


Figure A3.8: Scaled Schoenfeld residuals.

(a-e) Schoenfeld residuals plotted against time. In all cases, the proportional hazard assumption is supported by a non-significant association between residuals and time, thus the proportional hazards model can be assumed. The black solid line represents the best fitting line, while the grey shaded area shows the 95% CI for the regression. Test and plots modified from the *survminer* R package. **(a)** Residuals for lineage 2 and lineage 4 in Peru. **(b)** Residuals for lineage 2 and several sublineages of lineage 4 in Peru. **(c)** Residuals for sensitive and isoniazid mono-resistant background in Peru. **(d)** Residuals for lineage 2 and lineage 4 in Samara. **(e)** Residuals for sensitive and isoniazid mono-resistant background in Samara.

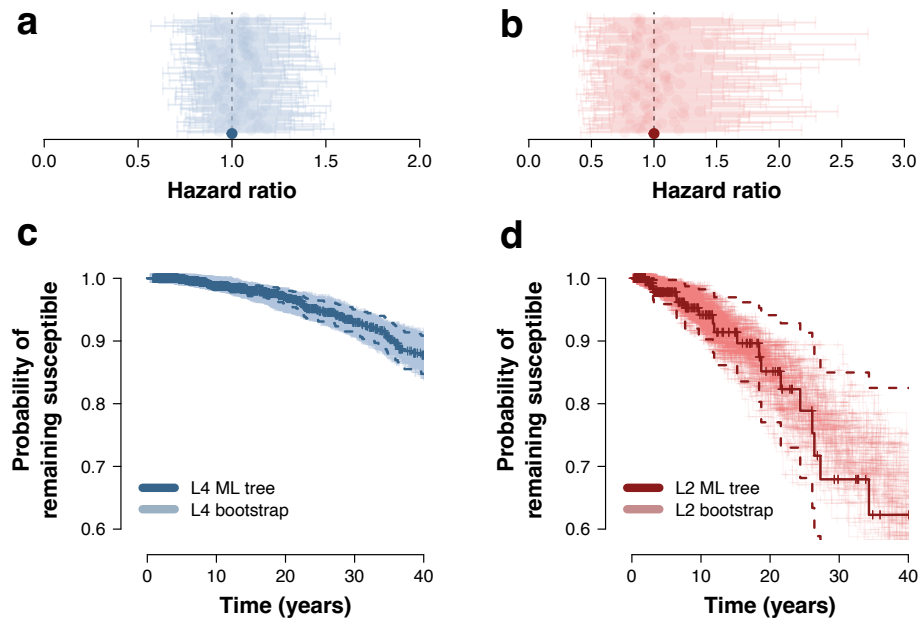


Figure A3.9: Hazard ratio and Kaplan-Meier curve of 100 phylogenetic bootstrap replicates.

100 phylogenetic bootstrap replicates were performed in parallel to the maximum likelihood phylogeny. Blue color represents the lineage 4 phylogeny while the red color shows the lineage 2 phylogeny. No statistical differences are found between the results obtained using the maximum likelihood tree and those using the phylogenetic bootstrap replicates. **(a, b)**, Hazard ratio for the maximum likelihood tree (dark point) and 100 phylogenetic bootstrap replicates (light colored points). Points and error bars represent the HR estimate and the 95% CI, respectively. **(c, d)**, Kaplan-Meier curve for the maximum likelihood tree (dark solid lines) with the 95% CI interval (dark dotted lines), and 100 phylogenetic bootstrap replicates (light colored lines).

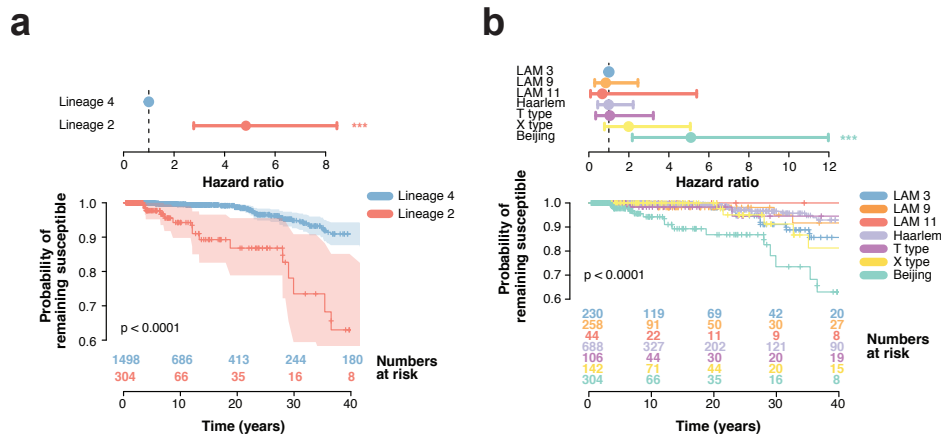


Figure A3.10: Hazard ratio and Kaplan-Meier curve for drug resistance acquisition in the population level dataset.

(a-b) Top: Hazard ratio (HR). Points and error bars represent the HR estimate and the 95% CI, respectively. The p-value for the HR was calculated using the likelihood ratio test. Bottom: Kaplan-Meier curve and numbers at risk. Y-axis represents the probability of remaining susceptible to any antibiotic, while the X-axis shows the time in years or the distance in branch length. Shaded areas show the 95% confidence interval. P-values for the Kaplan-Meier curves were calculated using the log-rank test. **(a)** HR of lineage 2 compared to lineage 4 in the Peruvian population level dataset (HR 4.84, 95% CI 2.78-8.45, Likelihood ratio test p-value = 2.7×10^{-8}) and the different Kaplan-Meier curve for lineage 2 and lineage 4 (log-rank test p-value = 7.9×10^{-10}). **(b)** HR between lineage 2 and the different sublineages of lineage 4 found in the Peruvian dataset (LAM9, LAM3, LAM11, Haarlem, X type and T type), using LAM3 as a reference (lineage 2 HR 5.1, 95% CI 2.17-11.9, Likelihood ratio test p-value = 1.8×10^{-4} , all other p-values > 0.2; Kaplan-Meier curve Log-rank test p-value = 3.02×10^{-7}). Statistical significance of the hazard ratio differences presented next to the CI bars (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$).

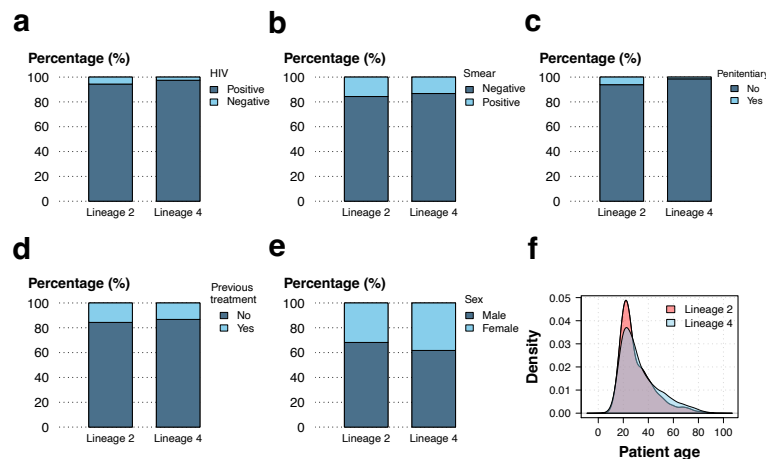


Figure A3.11: *Mycobacterium tuberculosis* lineage differences in non-bacterial factors.

Distribution of environmental factors in lineage 2 and lineage 4. (a), HIV status. (b), Smear positivity. (c), Imprisonment status. (d), Previous treatment with antitubercular drugs. (e), Patient sex. (f), Distribution of patients age.

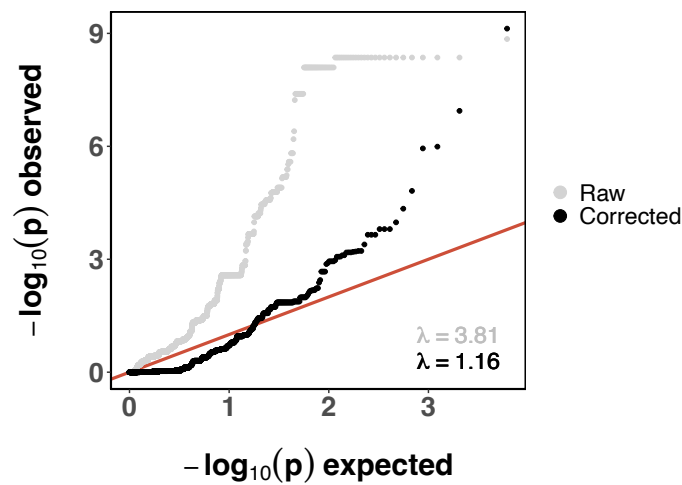


Figure A3.12: Population correction for Genome-Wide association study (GWAS). Quantile-quantile (QQ) for the raw p-values (grey), and the p-values corrected for population structure (black). Red line indicates the null hypothesis of uniformly distributed p-values. λ represents the genomic inflation factor.

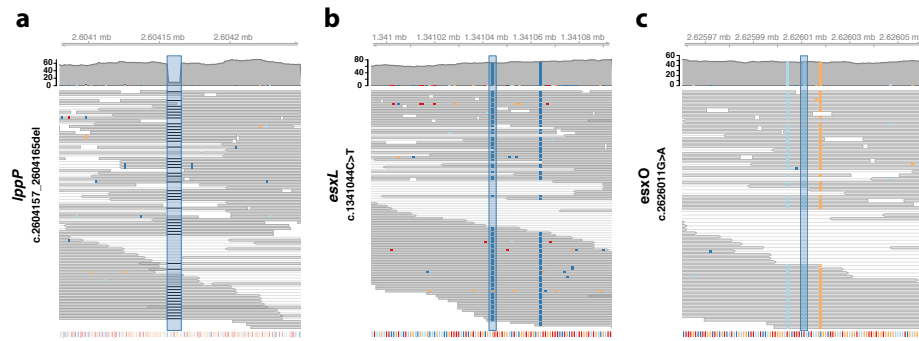


Figure A3.13: Short-read alignment around GWAS polymorphisms.

Coverage plot (top) and alignments from paired-end reads around GWAS hits. Alignments (middle panel) are represented as gray polygons with mismatches from the reference genome (bottom) indicated by different colors. Gaps are shown as black bars. Shaded area highlights the variant of interest. **(a)** *lppP* deletion. **(b)** SNP in the *esxL* gene, showing a T instead of the reference C. **(c)** Base mutation from G to A in the *esxO* gene.

Chapter 4

Table A4.1: Metrics used for phylogenetic tree comparison.

Abbreviation	Name	Rooted trees	Branch length	R function & package
KF	Kuhner and Felsenstein	No	Yes	KF.dist, phangorn
RF	Robinson Foulds	No	No	RF.dist, phangorn
wRF	Weighted Robinson Foulds	No	Yes	wRF.dist, phangorn
PS	Penny and Steel path	No	No	path.dist, phangorn
wPS	Penny and Steel path	No	Yes	path.dist, phangorn

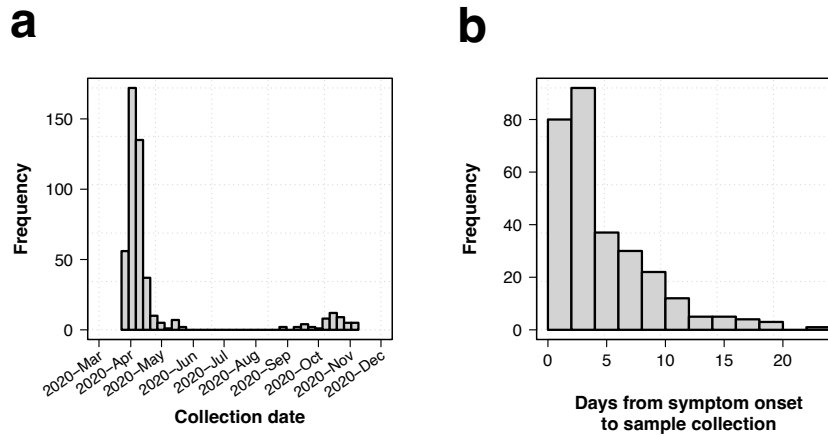


Figure A4.1: Collection date distribution and time from symptom and days from symptom onset.

(a) Distribution of collection dates. (b) Histogram of time from symptom onset to sample collection.

Table A4.2: Transition/transversion rates and base frequencies of the known simulated tree.

	A	Ag	Ca	Ga	C	Ta	Gc	G	Tc	Tg	Gt	Ac	Cg	T	At	Ct	Base freqs.
A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.22
Ag	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.01
Ca	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.01
Ga	0	200	0	-	-	-	-	-	-	-	-	-	-	-	-	-	0.01
C	0	0	1	0	-	-	-	-	-	-	-	-	-	-	-	-	0.22
Ta	0	0	0	0	0	-	-	-	-	-	-	-	-	-	-	-	0.01
Gc	0	0	0	0	0	0	-	-	-	-	-	-	-	-	-	-	0.01
G	0	0	0	1	0	0	1	-	-	-	-	-	-	-	-	-	0.22
Tc	0	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-	0.01
Tg	0	0	0	0	0	0	0	0	0	-	-	-	-	-	-	-	0.01
Gt	0	0	0	0	0	0	0	20	0	200	-	-	-	-	-	-	0.01
Ac	20	0	200	0	0	0	0	0	0	0	0	-	-	-	-	-	0.01
Cg	0	0	0	0	20	0	200	0	0	0	0	0	-	-	-	-	0.01
T	0	0	0	0	0	1	0	0	1	1	0	0	0	-	-	-	0.22
At	20	0	0	0	0	200	0	0	0	0	0	0	0	0	-	-	0.01
Ct	0	0	0	0	20	0	0	0	200	0	0	0	0	0	0	-	0.01

Table A4.3: Inferred transition/transversion rates and base frequencies when using the consensus sequence. Numbers show the average of 100 simulations.

	A	C	G	T	Base freqs
A	-	-	-	-	0.255
C	0.974	-	-	-	0.250
G	1.088	1.148	-	-	0.247
T	0.854	0.996	1	-	0.248

Table A4.4: Inferred transition/transversion rates and base frequencies when accounting for within-host diversity. Numbers show the average of 100 simulations.

	A	Ag	Ca	Ga	C	Ta	Gc	G	Tc	Tg	Gt	Ac	Cg	T	At	Ct	Base freqs.
A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.22
Ag	17.35	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.01
Ca	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.01
Ga	0	185.96	0	-	-	-	-	-	-	-	-	-	-	-	-	-	0.01
C	0	0	1.86	0	-	-	-	-	-	-	-	-	-	-	-	-	0.22
Ta	0	0	0	0	0	-	-	-	-	-	-	-	-	-	-	-	0.01
Gc	0	0	0	0	0	0	-	-	-	-	-	-	-	-	-	-	0.01
G	0	0	0	1.16	0	0	1.08	-	-	-	-	-	-	-	-	-	0.21
Tc	0	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-	0.01
Tg	0	0	0	0	0	0	0	0	0	-	-	-	-	-	-	-	0.01
Gt	0	0	0	0	0	0	0	21.64	0	196.56	-	-	-	-	-	-	0.01
Ac	16.51	0	183.53	0	0	0	0	0	0	0	0	-	-	-	-	-	0.01
Cg	0	0	0	0	21.20	0	200	0	0	0	0	0	-	-	-	-	0.01
T	0	0	0	0	0	0.98	0	0	1.05	1.26	0	0	0	0	-	-	0.21
At	17.70	0	0	0	0	175.58	0	0	0	0	0	0	0	0	0	-	0.01
Ct	0	0	0	0	20.09	0	0	0	188.86	0	0	0	0	0	0	0	0.01

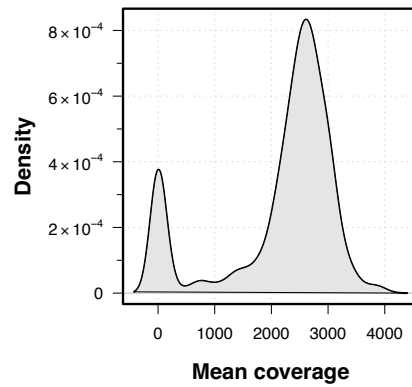


Figure A4.2: Sample mean coverage distribution.
Density distribution of mean coverage.

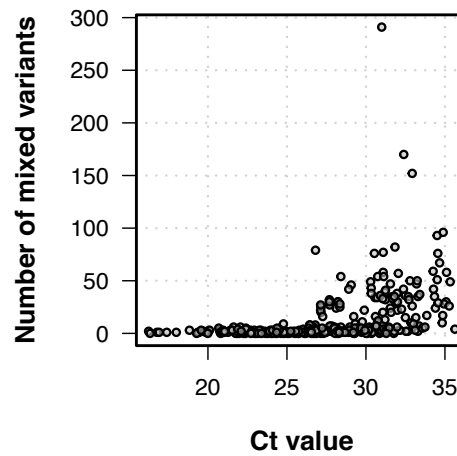


Figure A4.3: Number of low frequency variants and C_t value.
Higher C_t values were linked to a higher number of within-sample variation.

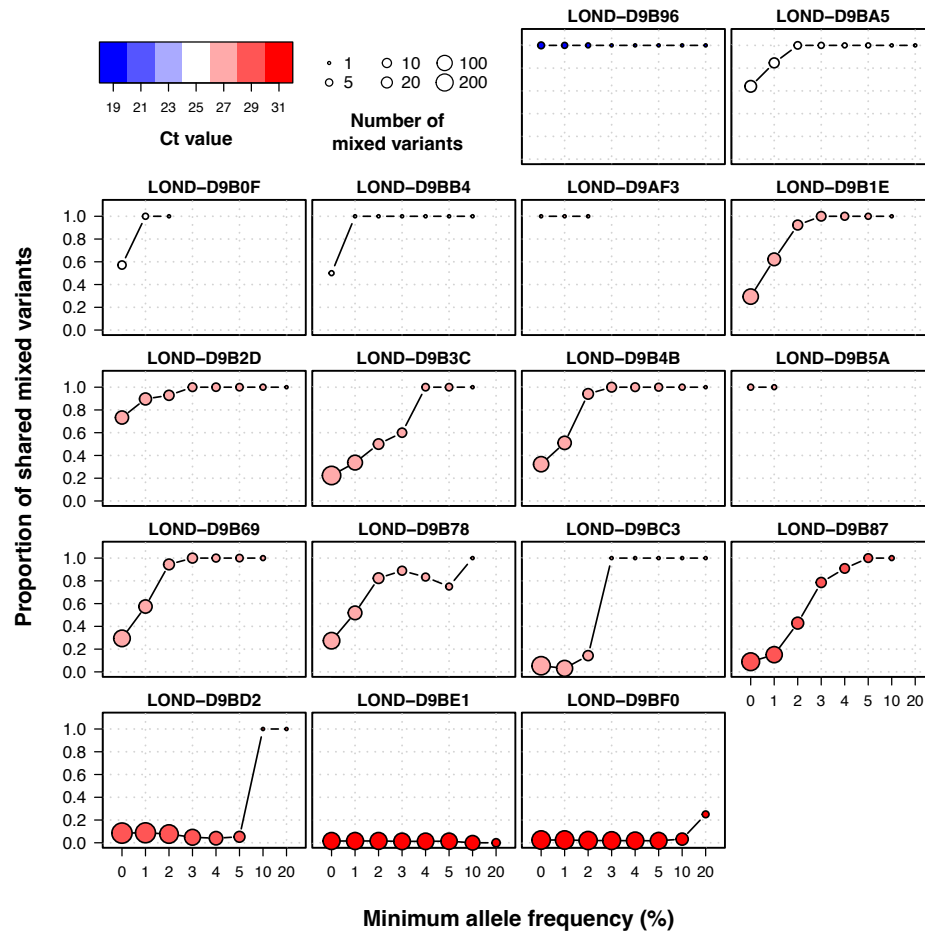


Figure A4.4: Proportion of shared mixed variants between duplicated samples using different filters of allele frequency.

Individual plots of shared within-host variants between technical duplicates using increasing thresholds of allele frequency. Colors represent C_t value, while the size of the point shows the total number of within-host variants between the two samples.

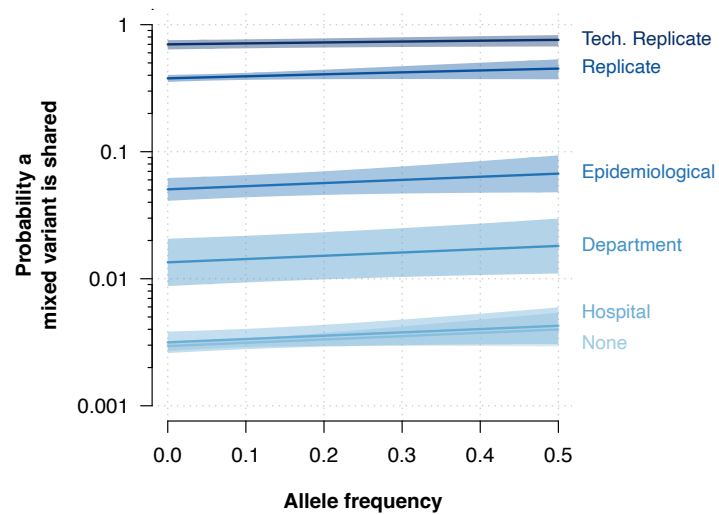


Figure A4.5: Probability that mixed variants are shared.

Probability that low frequency variants are shared inferred with a logistic model with allele frequency and epidemiological relationship as independent variable and whether a variant is shared or not as dependent variable. Y-axis in logarithmic scale for representation.

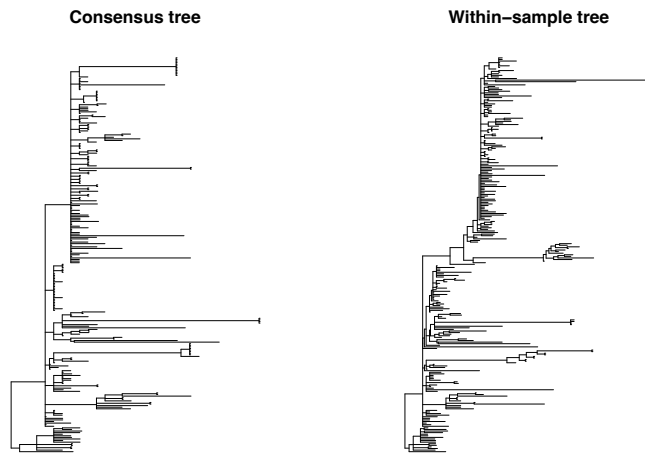


Figure A4.6: Phylogenetic trees for SARS-CoV-2.

SARS-CoV-2 phylogenetic trees inferred from consensus sequences (left) and an alignment with major and minor variant information (right).

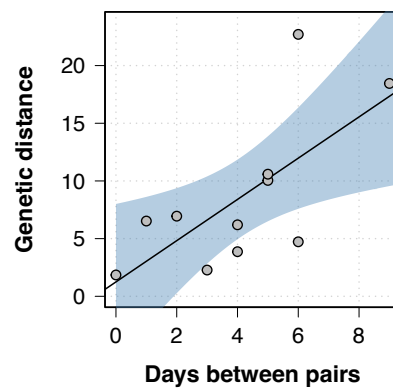


Figure A4.7: Genetic distance between longitudinal samples.

The genetic distance in the phylogenetic tree inferred using within-sample diversity increased as the between longitudinal samples progressed. Black line shows the best fit in a linear model, while the blue shaded area represents the 95% CI.

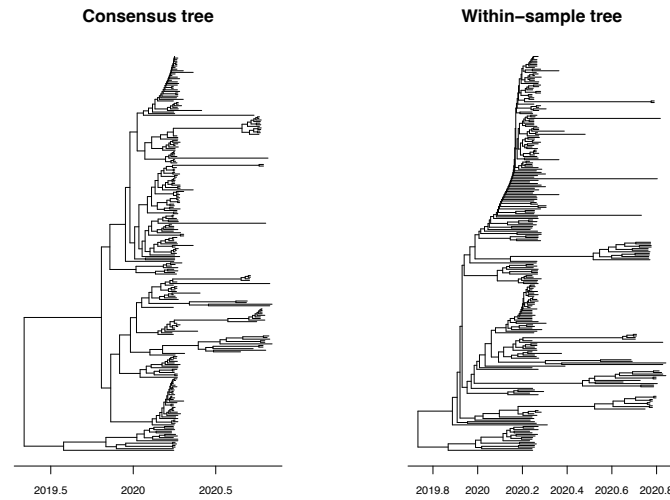


Figure A4.8: Time calibrated phylogenetic trees for SARS-CoV-2.

SARS-CoV-2 phylogenetic trees inferred from consensus sequences (left) and an alignment with major and minor variant information (right). Branch lengths are measured in years.

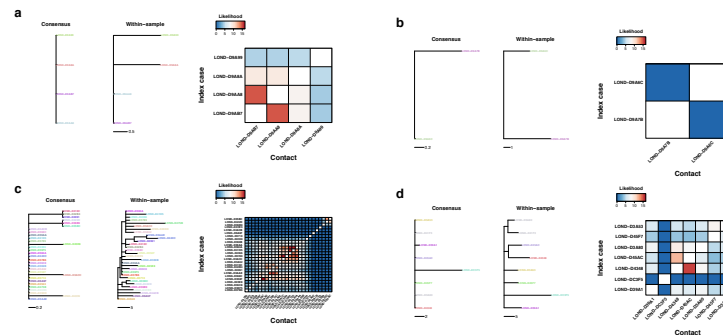


Figure A4.9: Phylogenetic and transmission for SARS-CoV-2 outbreaks.

a-d Phylogenies of SARS-CoV-2 outbreaks. The branch lengths are in units of substitutions per genome, and the scales are shown under the trees. Colors represent samples from the same individual. Samples with the same name are technical replicates. Left tree of each panel shows the phylogeny inferred with the consensus alignment. Right tree represents the phylogeny inferred using within-sample variation. Heatmap shows the likelihood of direct transmission for each pair of samples in a SEIR model of transmission. Vertical axis is the infector while the horizontal axis shows the infectee.