Mycobacterium tuberculosis genomics: The Next Generation

Anyansi, C.A.

DOI
10.4233/uuid:2276c5ba-1ec3-4ff7-b0fb-141175f4c76f

Publication date
2021

Document Version
Final published version

Citation (APA)
Anyansi, C. A. (2021). *Mycobacterium tuberculosis genomics: The Next Generation*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:2276c5ba-1ec3-4ff7-b0fb-141175f4c76f

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Mycobacterium tuberculosis genomics: the next generation

Christine Anyansi

Mycobacterium tuberculosis genomics: the next generation

Dissertation

For the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof.dr.ir.T.H.J.J. van der Hagen
chair of the board for doctorates
to be defended publically on
Thursday 9 December at 12:30 hours

By

Christine ANYANSI

Master of science in Bioinformatics, Vrije Universiteit, Netherlands
Born in Ogidi, Nigeria

This dissertation has been approved by the promoters.

Composition of the doctoral committee:

Rector Magnificus,	chairman
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology, promotor
Dr T.E.P.M.F Abeel,	Delft University of Technology, copromotor

Independent members:

Prof.dr. A. Van Rie	U. of Antwerp, Belgium
Prof.dr. V. van Noort	Leiden U.
Prof.dr. P.A.S. Daran-Lapujade	TNW
Prof.dr. A.M. Dogterom	TNW
Prof.dr. N.H. Dekker	TNW

TU Delft, Delft Global

Keywords: tuberculosis, genomics, bioinformatics, global health, sequencing
Printed by: ProefschriftMaken

ISBN 978-94-6423-586-9
An electronic version of this dissertation is available at
http://repository.tudelft.nl

Outline

Contents

Summary

The greatest threat to mankind lies within the microscopic organisms we co-exist with. Indeed, wars have raged between humans and the pathogens invading human bodies for all the history of mankind. The unprecedented scale and impact of the Covid-19 pandemic demonstrates again the power of the microscopic to bring humanity to a collective heel, much like smallpox, plague and others have shown before.

Tuberculosis (TB), also known as consumption, the white plague, phthisis; may very well be the oldest infectious disease known to man. As such, its history has been extensively studied. Traces of mycobacterium tuberculosis (MTB) the causative bacterial pathogen for TB, have been found in multiple skeletal remains, dating as far back as 9000 years ago. Remarkably, evidence of TB lesions have also been found from homo-erectus fossils in Turkey dating back 50,000 years!

After millennia of sparring with MTB, our fight with this pathogen is still far from over. TB is currently the single most deadly infectious disease in the world, taking the lives of millions each year and infecting a quarter of humanity. The rapid rise of antibiotic resistance is an effective counter move by MTB and poses a serious and real public-health threat. For such an old and ancient pathogen, it is perplexing that it still remains at large. What makes TB such a horrible and deadly disease? How can we best fight TB? What treatment regimens are most effective? How does TB spread? How can we track outbreaks? Why are some types of TB more infectious than others? How should we detect TB in humans? What drives antibiotic resistance? How can the spread of antibiotic resistance be halted?

In short: How can we win this war against TB?

The turn of this century has brought with it a strong playing card – whole genome sequencing (WGS). The ability to analyze MTB WGS data en masse has resulted in a plethora of unprecedented insights, discoveries and innovations. With WGS we can upgrade the way we determine policies and make decisions for our tactical strategy against TB in a data driven approach.

This thesis represents one small step in the battle with TB, where we apply WGS approaches to explore different topics of TB research. Here we offer the global community a method to diagnose complex TB infections consisting of multiple distinct strains. We show that this functionality is necessary and has been overlooked by the TB community in research studies, which might have contributed to poor treatment outcomes. As the increase in samples resistant to multiple antibiotics is a pressing challenge, we explore global trends in antibiotic resistance evolution and have identified a particular order of resistance acquisition for 6 anti-tubercular drugs. This thesis also provides complete assemblies of 18 MTB genomes spanning 7 lineages that was used to analyze MTB's largest and least studied gene family.

Overall, this thesis intends to shed light on lesser explored topics in TB research, while providing insights that could be built upon to further our knowledge of MTB and ultimately save some lives in the process.

1 Introduction

1.1 About TB

The bacteria *Mycobacterium tuberculosis* (MTB) has been a pathogen of world-wide concern since its first discovery in 1882 by the scientist Robert Koch[1]. It has caused an estimated 10 million new infections, and 1.4 million deaths in the year 2019 – making it the top killer from an infectious disease around the world [2]. The End TB Strategy formulated by the WHO calls for an 80% reduction of new tuberculosis (TB) cases and a 90% reduction of TB deaths by 2030 [3]. To achieve these goals the global community has banded together to find innovative methods and reduce the global TB burden that disproportionately effects impoverished societies.

The pathogenesis of MTB has been well studied throughout the years, giving us insight into how the pathogen has successfully been able to penetrate through all time periods of civilization. MTB is an airborne pathogen transmitted between patients in the form of aerosolized particles emitted while coughing [4,5]. Not all patients harboring the pathogen are characterized with an active infection. Most individuals (95%) first exposed to TB remain asymptomatic – what is referred to as a latent infection. This occurs when the immune system successfully manages to contain the invading pathogen in granulomas [6]. Because the bacteria is only contained and not eradicated, TB patients with latent infection either remain asymptomatic for the rest of their lives, or develop an active infection at a later time point [6,7]. Once an active TB infection develops, the patient may transmit MTB to others and experience the typical symptoms such as cough, chest pain, fatigue, and fever. Without proper treatment, the patient faces long-term lung damage [8] and in unfortunate cases: death.

1.2 Tuberculosis Diagnosis

Early diagnosis of TB is fundamental in effectively managing the disease as this leads to timely treatment and reduces future transmission [9,10]. General first steps can involve a chest X-ray [11], CT-scan [12] or a tuberculin skin test [13]. These, relatively quick methods however do not provide a definitive diagnosis, which has traditionally been achieved via cultured sputum [14]. Diagnosis via culture is a time intensive process and can only detect TB in sputum samples containing a sufficient amount of the pathogen [15]. Developing quicker and more sensitive tests to diagnose TB has therefore been a field of active research.

1.3 Tuberculosis Treatment

Currently, TB is controlled either with a preventative approach, i.e. through a vaccine, or a treatment based approach through chemotherapy (drugs). The only licensed TB vaccine, Bacillus Calmette–Guérin (BCG) has been in use for over a century but is applicable only against childhood TB and offers nearly no protection for adult TB [16]. Despite administering the vaccine to children in endemic countries, high rates of TB infection for adults continue to persist [16]. Developing new vaccines for wide spread protection against TB is a top priority for the WHO and requires advanced research into the genetic mechanisms of TB to find putative vaccine targets.

Drug	Known genes with mutations that confer resistance	Drug type
rifampicin (RIF)	*rpoB*	rifampycin, first-line
isoniazid (INH)	*katG*, *inhA**, *FabG**, *kasA*	analagous to eth, first-line
pyrazinamide (PYR)	*pncA*, *clpC1*, *panD*	pyrazine, first-line
ethambutol (EMB)	*embB*, *embAB*, *embC*	bacteriostatic, first-line
streptomycin (STR)	*rpsL*, *gidB*, *murA*	aminoglycide, first-line
ofloxacin (OFL)	*gyrA*, *gyrB*	fluoroquilone, second-line
kanamycin (KMC)	*eis*, *murA*	aminoglycide, second-line
ethionamide (ETH)	*ethA*, *inhA**, *fabG**	analogous to INH, second-line
bedaquiline (BDQ)	*pepQ*, *Rv0678*, *mmpL5*, *mmpS5*, *atpE*	diarylquinolines, second-line
linezolid (LZD)	*rplC*, *rrl*	linezolid, second-line

Table 1.1 Drugs given to treat TB, along with genes that have been identified to be associated with antibiotic resistance. The type of drug and whether it is a first line (given first) or second line (given after first line drugs have failed) are also shown.

The most effective method to control active TB infections is via the administration of several antibiotic drugs (Table 1.1) over the course of six to nine months. The specific regimen can differ by country, but the standard treatment involves a combination of the drugs isoniazid, rifampicin, pyrazinamide and ethambutol for two months, followed by only isoniazid and rifampicin for an additional two months [17].

1.4 Tuberculosis Drug Resistance

The emergence and spread of antibiotic resistance has led to poor treatment outcomes with global treatment success rates of 57% [2] making the TB epidemic particularly difficult to control. Poor treatment and poor containment of patient to patient spread has led to continued proliferation of multi-drug resistant TB (MDR-TB) a type of TB resistant to two of the first line drugs: rifampicin and isoniazid [18]. Six percent of MDR-TB cases are estimated to be extensively drug resistant TB (XDR-TB), a type of TB resistant to isoniazid, rifampicin, any fluoroquinolone, and an injectable second line drug (amikacin, kanamycin or capreomycin) according to the WHO definition of 2019. A cure for XDR-TB is only possible for 30-50% of cases [19].

1.5 Drug Resistance Diagnosis

Timely detection of antibiotic resistant TB is crucial for prescribing a more successful treatment regimen for patients. With this in mind, drug susceptibility tests (DSTs) are typically performed to diagnose presence of drug resistance within patients. Culture-based DSTs defines resistance when more than 5% of MTB population can grow in a medium containing a ‘critical concentration’ of a specified drug [20]. A sample unable to grow at the drug levels are therefore labelled as susceptible, as they are susceptible to treatment with the drug. This method of diagnosis comes with many problems. Concentration levels were not based on pharmacokinetics but rather comparing wild type strains, which were never exposed to drugs or sensitive, to non-wild type

strains (that have been exposed to drugs) [21] and levels fluctuate depending on the type of medium used (i.e. egg based vs synthetic)[20]. Finally, the time to results is long due to the slow growth rate of MTB, and can take 6 to 8 weeks [22].

In contrast to many other bacteria, most of drug resistance observed in MTB is due to single nucleotide polymorphisms (SNPs) [23] making molecular based techniques an attractive option to quickly diagnose antibiotic resistance. Amongst the plethora of molecular based assays developed to diagnose antibiotic resistance in TB, WHO endorses the Gene Xpert Ultra assay, the newest version of the previously used Gene XPERT MTB/RIF assay as an initial susceptibility test instead of culture [24]. The Ultra is a PCR based test that probes for mutations in *rpoB* genes, conferring resistance to rifampicin, with a very short turn-around time. Although sufficient for detecting rifampicin resistance, current molecular based assays are not only limited in the breadth of drugs they can survey but require in depth understanding of the genetic basis of TB drug resistance in order to select suitable SNPs associated with resistance.

1.6 Tuberculosis population structure

MTB belongs to a highly related group of mycobacteria (99.9% nucleotide identity) that cause TB in both human and animals, referred to as the mycobacterium tuberculosis complex (MTBC) [25]. Several lineages primarily infect animal hosts, such as *M. bovis* (cows), and *M. microti* (rodents) and it was originally believed that TB originated in animals [26]. However, thanks to genome sequencing and comparative genomics it was discovered that both animal and human lineages shared a common ancestor [27–29]. The seven human adapted lineages of MTBC consists of *M. tuberculosis sensu stricto* (Lineages 1-4, Lineage 7), and *M. africanum* (Lineages 5-6). Additionally, a new human adapted lineage of MTBC has been identified (Lineage 8) with a branch point before the MRCA of the rest of the human lineages [30]. The human adapted MTB lineages vary with respect to their geographical spread as some lineages are present across a wide range of countries while others are restricted to a particular geographic area (Figure 1.1). The naming of the lineages also follow their geographical patterns. Lineage 1: East African Indian or EAI lineage, Lineage 2: East-asian or Beijing lineage, Lineage 3: Central Asian Strain (CAS) lineage, Lineage 4: Euro American lineage (occurring globally), Lineage 5: *M. africanum* West Africa 1, Lineage 6: *M. africanum* West Africa 2, Lineage 7 occurs only in Ethiopia and the new Lineage 8 was discovered in Rwanda and Uganda.

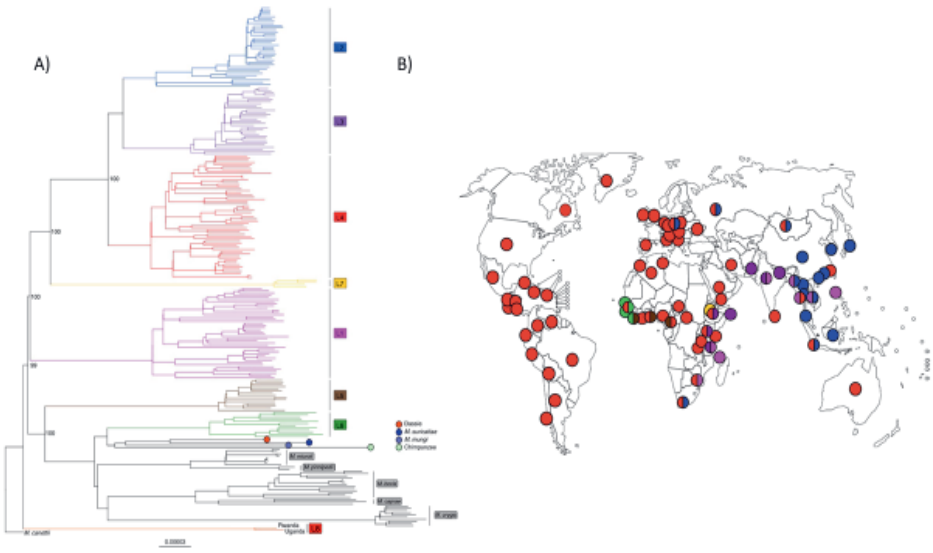


Figure 1.1 A) Phylogeny of several lineages that comprise the MTBC complex using 241 genomes from Ngaboniza et al [31]. Tree was made using maximum likelihood and rooted using M. canetti as an outgroup. B) Geographical distribution of MTBC coloured by the main lineage in the region according to the scheme in panel A. Sourced from Brites et al [25].

Lineages of MTB not only vary geographically but also exhibit phenotypic differences [32]. Lineages differ in their transmissibility with lineage 2 strains commonly reported to have increased virulence levels [33], and relatively quick progression from latent to active infection [34]. Lineages differ in their interactions with the human immune system, with lineages 1 and 2 inducing a greater response from macrophages in animal models [35]. Additionally, studies have shown lineage specific patterns in the acquisition and occurrence of drug resistance [36,37].

1.7 Molecular methods to track tuberculosis

Typing methods or methods that differentiate between strains/lineages/genotypes of TB are needed in order to understand the epidemiology of MTB, provide a mechanism to track outbreaks, and define source of infections in transmission events. Furthermore occurrences of infections caused by multiple MTB strains, either via within host microevolution or transmission events have steadily been reported in literature and have been linked to treatment outcomes [38]. Molecular genotyping methods were developed with these goals in mind. The major methods include IS*6100*-based restriction fragment length polymorphism (IS*6100*-RFLP), polymorphic GC-rich repetitive sequence RFLP (PGRS-RFLP), spoligotyping, and mycobacterial interspersed repetitive units variable number of tandem repeats (MIRU-VNTR).

IS*6100*-RFLP has been in use since the early 1990s and was the former gold standard for strain genotyping [39,40]. It determines genotypes based on the number of copies of the IS*6100* mobile

element which can vary between 0 and 20. Although once widely used, this technique has low discriminatory ability for MTB strains with less than six IS*6100* bands, requires culturing due to its high DNA volume requirement (2-3 ug), and is difficult to compare results across labs as visual inspection is necessary. PRGS-RFLP was developed as a supplementary method to gain higher resolution for strains with less than six IS*6100* bands based on the PGRS region of the genome. However it still suffers from similar pitfalls as IS*6100*-RFLP with the additional complexity to read/interpret as there are greater number of bands[40].

Spoligotyping was also developed for MTB isolates with less than six IS*6100* bands [41]. It determines genotypes based on PCR amplification of direct variable repeat regions within the genome that are lineage/strain specific [42]. As opposed to IS*6100*-RFLP, it is more reproducible and requires much less DNA [40]. However it is not as specific as IS*6100*-RFLP and MIRU-VNTR typing.

MIRU-VNTR has higher discriminatory power and is more reproducible than both spoligotyping and IS*6100*-RFLP [43,44] making it the gold standard of conventional molecular genotyping methods. It determines genotypes based on PCR amplification of VNTRs within 24 MIRU regions of the genome.

Overall molecular genotyping techniques differ with respect to their reproducibility, discriminatory power and technical requirements. They have been shown to sufficiently discriminate between lineages but do poorly for intra lineage strain discrimination [44].

1.8 Whole genome sequencing and tuberculosis

1.8.1 Background TB genomics

The genomic revolution has been paramount in shaping our understanding of TB. Since the first full assembly of the MTB genome in 1998 [45], whole genome sequencing (WGS) has been applied to almost every facet of TB research and clinical practice. Indeed, England became in 2017 the first country to introduce WGS in its routine clinical pipelines for TB management [46]. Through WGS, many of the time consuming steps within the clinical management of TB can be reduced compared to traditional methods, such as diagnosis, strain classification, and drug susceptibility testing – all of which can be conducted within a single WGS analysis. Additionally WGS offers unprecedented insights into the population structure, evolution and mutation rates of MTB genotypes, far greater than traditional molecular approaches. With the decreasing costs of WGS, the future of WGS-based MTB analysis has plenty of room for expansion.

The first MTB assembly was constructed using shotgun sequencing with cosmids and BAC clones. Albeit highly accurate shotgun sequencing was a time intensive process. Since then, advances in high throughput sequencing have drastically sped up analysis and Illumina based short-read sequencing has become the dominant technology to perform genomic analysis of MTB. Typically, in analysis of MTB with short-reads, approximately 10% of the genome is discarded due to long GC-rich repetitive stretches of DNA within the PE/PPE gene families. It is speculated that these gene families play an important role in pathogenicity [47] and virulence[48], making it important to find suitable ways to include these families in standard MTB studies. In the past few years third generation sequencing represented by PacBio and Oxford Nanopore technologies have facilitated

sequencing of long reads of lengths only limited by the length of the DNA molecule. Long read sequencing trumps short-reads due to the size of the reads, capable of spanning difficult to sequence regions of genomes, however it suffers from a higher error rates in base-called reads (3% vs 0.1% of Illumina [49,50]). Although Illumina is currently dominating sequencing analysis of MTB, it is likely that further improvements and accessibility of Nanopore or PacBio technologies will lead to increasing adoption of long-read sequencing.

1.8.2 WGS for identifying strains

Epidemiological studies based on molecular genotyping have elucidated the importance of identifying individual strains of MTB in order to track outbreaks, study evolutionary history, distinguish between relapse and reinfection. Techniques to identify individual strains can also be used to discover links between cases, and explore within host heterogeneity [51–53] . WGS is an alternative to using molecular techniques which examine only a subset of the genome. WGS can capture full scale of genetic variation and provides accurate information for delineating strains in a reproducible manner [54]. WGS has been used to identify direct transmission events between source and target based on SNV distance [54]. SNV distance refers to the number of variations that differ between two strains. If there are only a few number of sites that strains differ it would be difficult to tell which one is which. Therefore the higher the number of variation the easier it is to differentiate strains. Molecular methods are not capable of identifying with such low resolution and are also incapable of quantifying distance between two strains, while this is possible with WGS. WGS has been widely applied to tracking transmission outbreaks based on SNPs [55–59]. More sophisticated algorithms have been developed to identify and characterize strains of a particular lineage within isolates, borrowing off of principles from the metagenomics field.

1.8.3 WGS impact in antibiotic resistance

As MTB resistance emerges due to the variation within the genome (i.e. from SNPs, insertion, deletions), methods to detect such genomic variations can fundamentally improve clinical management of drug resistant TB cases. Therefore a patient's observed SNP profile can be used to predict which drugs would be clinically effective and also used as a tool to monitor the effectiveness of treatment, by checking for the emergence of any new resistance informative SNPs that could arise due to a reinfection from a different strain or through improper treatment regimes. Currently, WGS-based drug susceptibility testing still requires a culture step of the bacteria. This leads to resistance results being biased to that of the sampled culture colony.

Numerous studies have characterized genomic variation that are indicative of resistance [60], however this is still a work in process as many causes of resistance have yet to be identified. The WHO maintains a catalogue of resistance genes for TB [61], which is the most up to date list of resistances. Through WGS, the occurrence of antibiotic mutations arising via compensatory or convergent means have been revealed. Compensatory mutations refer to mutations that do not directly lead to resistance but lead to increased fitness in the presence of resistance conferring mutations, thereby offsetting any loss of fitness presented by the antibiotic resistance mutation. Compensatory mutations subvert the expectation that antibiotic resistance comes with loss of fitness for the bacteria [54,62]. Mutations arising via convergent evolution have evolved

independently several times across lineages [63]. Understanding the evolution and underlying process of antibiotic resistance is a complex task as there are different pathways involved that can interact with each other. However, there is an increasing amount of evidence that accumulation of resistance mutation occurs in a predictable manner[64–66]. Understanding the evolution of resistance conferring mutations can provide valuable insight into the preventing DR-TB and disrupting the emergence of MDR/XDR-TB.

Several online software tools incorporate databases with associations between genetic variations and susceptibility. PhyResSe [67] and TB Profiler [68] are both online tools that process fastq files to output drug resistance profiles, while MyKrobe Predictor [69] and KvarQ [70] are downloadable executables that perform the same function. PolyTB [71] is an online tool to facilitate searching of SNPs within the MTB genes to aid in resistant mutation detection. The ReSeqTB platform [72] contained its own variant calling pipeline in addition to a database of genomic data paired with phenotypical meta data such as culture-based DST results and clinical outcome. The TBDream database [73] contained information about frequency of mutations in resistant vs susceptible isolate but is no longer accessible online. Mycoresistance offers a manually curated database of compounds used to treat TB with their associated resistances pulled from hundreds of studies [74]. Previous studies have shown that genotypic predictions for drug resistance are in line with the underlying resistance phenotype[75,76]. Extensive work has been performed to illustrate the applications of WGS on drug susceptibility testing and its ability to reduce the time and price required for results [77].

1.8.4 Heteroresistance

Heteroresistance, referring to the co-existence of susceptible and resistant variations, complicates management of patients with drug resistant TB. This phenomena is caused by the infection of multiple MTB strains or through within host microevolution and has been observed in at least 5% of drug resistant TB cases [78]. Culture of MTB is not always able to detect heteroresistance [79] which could lead to inappropriate treatment. WGS has facilitated the quantification of heterogeneity at drug resistant loci of patients [80,81] with the abundance of minority variants fluctuating over the infection until the mutation becomes fixed [80,82]. The ability to identify heteroresistance and heterogeneity within infections could have a positive impact on patient outcomes.

1.9 Research Goals

With this thesis, I aim to apply WGS technologies to aid in the diagnosis, monitoring, and research of TB. I focus on three main areas that can be improved with WGS in TB research.

I first sought to replace the traditional typing tools that have been used to identify different strains/genotypes in MTB samples. These tools had been used to classify the presence of mixed TB infections; a use case that they were not originally intended for. Therefore, I created a method to identify individual strains present in TB WGS data (Chapter 2). In doing so I add to the broader field of metagenomics style tools, which were at the time not suitable for TB data.

As there were not any established tools to classify mixed infections in WGS data, the impact of this phenomena was over looked in research studies. The incidence of mixed infections, it's global dispersion, and their relation to antibiotic resistance are all topics that were only speculated at, not properly researched. Additionally, the research studies for other TB topics that used WGS data tended to overlook the possibility of heterogeneity contaminating their data. Therefore, I also wanted to understand how the presence of mixed infections could influence research studies, play a role in antibiotic resistance, and effect TB's global distribution patterns (Chapter 3).

The method in which antibiotic resistance is acquired is useful for making responsible decisions in the clinic and for deeper understanding of the bacteria. Though we know a lot about which mutations lead to antibiotic resistance for particular drugs, it's still not extremely clear the order in which these mutations evolve for all but the top 2 mutations. I looked into providing more insight into the arisal of antibiotic mutations using a large database of WGS isolates (Chapter 4).

As mentioned previously, the PE/PPE gene families representing 10% of the MTB genome, is typically discarded during routine bioinformatics analysis pipelines due to the presence of long repeat regions high in GC content. Though, these regions were speculated to play an important role for MTB, little is known about their functional variation across different MTB lineages as short-reads could not fully span the gene regions. Using long reads, we were able to completely sequence a representative set of MTB genomes. With them we aimed to provide more insight into this lesser studied gene families of TB (Chapter 4).

Metagenomics methods have broad applicability to fields also outside of TB and the their ability to perform strain-level resolution in bacterial populations has proved particularly challenging. Different tools have been developed for different use case scenarios, making finding the appropriate tool for a particular research goal quite cumbersome. As research aims steadily move towards more granular, fine grained insights at the strain level, I also sought to provide an overview of the tools developed to offer strain level insights for bacteria. Here I was also particularly interested in revealing what were the exact challenges such tools face and give a review on the state of the art algorithms (Chapter 5).

1.10 Thesis outline

This thesis aims to tackle unsolved issues within TB research using WGS techniques. Chapter 2 presents a method to conduct strain level classification of WGS TB data, particularly with the focus of detecting isolates containing a mixture of strains (mixed infections). Chapter 3 is a large scale study of the current state of TB WGS data from a meta-analysis of 50k isolates. It aims to reveal unidentified mixed infections present in previous research studies and explores the impact of this overlooked problem in current WGS applications for TB. Chapter 4 uses the TB data accumulated from Chapter 3 to explore evolutionary patterns of resistance acquisition in TB. In chapter 5 we apply long read sequencing technologies to provide the first complete set of TB assemblies spanning all lineages. We use these assemblies to study the elusive PE/PPE regions of TB. Lastly chapter 6 is focused on providing an overview of strain level classification techniques of bacterial genomes.

1.11 References

1. Koch R. Die Ätiologie der Tuberkulose (Nach einem in der physiologischen Gesellschaft zu Berlin am 24. März gehaltenem Vortrage). Berliner Klin. Wochenschrift 1882;

2. WHO. Global Tuberculosis Report 2020. J. Chem. Inf. Model. 2020;

3. World Health Organization. WHO | WHO End TB Strategy. World Heal. Organ. 2015;

4. Fennelly KP, Martyny JW, Fulton KE, et al. Cough-generated Aerosols of Mycobacterium tuberculosis: A New Method to Study Infectiousness. Am. J. Respir. Crit. Care Med. 2004;

5. Fennelly KP, Jones-López EC. Quantity and quality of inhaled dose predicts immunopathology in tuberculosis. Front. Immunol. 2015;

6. Kaufmann SHE. Protection against tuberculosis: Cytokines, T cells, and macrophages. Ann. Rheum. Dis. 2002;

7. Shiloh MU. Mechanisms of mycobacterial transmission: How does Mycobacterium tuberculosis enter and escape from the human host. Future Microbiol. 2016;

8. Ravimohan S, Kornfeld H, Weissman D, et al. Tuberculosis and lung damage: From epidemiology to pathophysiology. Eur. Respir. Rev. 2018;

9. Storla DG, Yimer S, Bjune GA. A systematic review of delay in the diagnosis and treatment of tuberculosis. BMC Public Health 2008;

10. Sreeramareddy CT, Panduru K V., Menten J, et al. Time delays in diagnosis of pulmonary tuberculosis: A systematic review of literature. BMC Infect. Dis. 2009;

11. Skoura E, Zumla A, Bomanji J. Imaging in tuberculosis. Int. J. Infect. Dis. 2015;

12. Pelletier-Galarneau M, Martineau P, Zuckier LS, et al. 18F-FDG-PET/CT Imaging of Thoracic and Extrathoracic Tuberculosis in Children. Semin. Nucl. Med. 2017;

13. CDC. TB elimination tuberculin skin testing. Cent. Dis. Control 2011;

14. Datta S, Shah L, Gilman RH, et al. Comparison of sputum collection methods for tuberculosis diagnosis: a systematic review and pairwise and network meta-analysis. Lancet Glob. Heal. 2017;

15. Mase SR, Ramsay A, Ng V, et al. Yield of serial sputum specimen examinations in the diagnosis of pulmonary tuberculosis: A systematic review. Int. J. Tuberc. Lung Dis. 2007;

16. Hatherill M, White RG, Hawn TR. Clinical Development of New TB Vaccines: Recent Advances and Next Steps. Front. Microbiol. 2020;

17. Nahid P, Dorman SE, Alipanah N, et al. Executive Summary: Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America Clinical Practice Guidelines: Treatment of Drug-Susceptible Tuberculosis. Clin. Infect. Dis. 2016;

18. WHO. Tuberculosis: Multidrug-resistant tuberculosis (MDR-TB). 2018;

19. CDC. Fact Sheets | Drug-Resistant TB | Extensively Drug-Resistant Tuberculosis (XDR TB) | TB | CDC. 2016;

20. Schön T, Miotto P, Köser CU, et al. Mycobacterium tuberculosis drug-resistance testing: challenges, recent developments and perspectives. Clin. Microbiol. Infect. 2017;

21. Canetti G. Present aspects of bacterial resistance in tuberculosis. Am. Rev. Respir. Dis. 1965;

22. CANETTI G, FROMAN S, GROSSET J, et al. MYCOBACTERIA: LABORATORY METHODS FOR TESTING DRUG SENSITIVITY AND RESISTANCE. Bull. World Health Organ. 1963;

23. Zhang Y, Yew WW. Mechanisms of drug resistance in Mycobacterium tuberculosis: Update 2015. Int. J. Tuberc. Lung Dis. 2015;

24. Hasnain SE, Ehtesham NZ, Grover S. Mycobacterium tuberculosis: Molecular infection biology, pathogenesis, diagnostics and new interventions. Mycobacterium Tuberc. Mol. Infect. Biol. Pathog. Diagnostics New Interv. 2019;

25. Brites D, Gagneux S. the nature and evolution of genomic diversity in the mycobacterium tuberculosis Complex. Adv. Exp. Med. Biol. 2017;

26. Stead WW, Eisenach KD, Cave MD, et al. When did Mycobacterium tuberculosis infection first occur in the New World? An important question with public health implications. Am. J. Respir. Crit. Care Med. 1995;

27. Comas I, Coscolla M, Luo T, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat. Genet. 2013; 45:1176–1182

28. Brosch R, Gordon S V., Marmiesse M, et al. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc. Natl. Acad. Sci. U. S. A. 2002;

29. Comas I, Gagneux S. The past and future of tuberculosis research. PLoS Pathog. 2009;

30. Ngabonziza JCS, Loiseau C, Marceau M, et al. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. bioRxiv 2020;

31. Ngabonziza JCS, Loiseau C, Marceau M, et al. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. Nat. Commun. 2020;

32. Thwaites G, Caws M, Chau TTH, et al. Relationship between Mycobacterium tuberculosis genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. J. Clin. Microbiol. 2008;

33. Tram TTB, Nhung HN, Vijay S, et al. Virulence of Mycobacterium tuberculosis Clinical Isolates Is Associated With Sputum Pre-treatment Bacterial Load, Lineage, Survival in Macrophages, and Cytokine Response. Front. Cell. Infect. Microbiol. 2018;

34. de Jong BC, Hill PC, Aiken A, et al. Progression to Active Tuberculosis, but Not Transmission, Varies by Mycobacterium tuberculosis Lineage in The Gambia . J. Infect. Dis. 2008;

35. Krishnan N, Malaga W, Constant P, et al. Mycobacterium tuberculosis lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. PLoS One 2011;

36. Oppong YEA, Phelan J, Perdigão J, et al. Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. BMC Genomics 2019;

37. Ford CB, Shah RR, Maeda MK, et al. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat. Genet. 2013;

38. Cohen T, Chindelevitch L, Misra R, et al. Within-host heterogeneity of mycobacterium

tuberculosis infection is associated with poor early treatment response: A prospective cohort study. J. Infect. Dis. 2016; 213:1796–1799

39. Houben RMGJ, Glynn JR. A systematic review and meta-analysis of molecular epidemiological studies of tuberculosis: Development of a new tool to aid interpretation. Trop. Med. Int. Heal. 2009;

40. Kato-Maeda M, Metcalfe JZ, Flores L. Genotyping of Mycobacterium tuberculosis: Application in epidemiologic studies. Future Microbiol. 2011;

41. Flores L, Jarlsberg LG, Kim EY, et al. Comparison of restriction fragment length polymorphism with the polymorphic guanine-cytosine-rich sequence and spoligotyping for differentiation of Mycobacterium tuberculosis isolates with five or fewer copies of IS6110. J. Clin. Microbiol. 2010;

42. Zheng C, Li S, Luo Z, et al. Mixed infections and rifampin heteroresistance among mycobacterium tuberculosis clinical isolates. J. Clin. Microbiol. 2015; 53:2138–2147

43. Weniger T, Krawczyk J, Supply P, et al. MIRU-VNTRplus: A web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria. Nucleic Acids Res. 2010;

44. Comas I, Homolka S, Niemann S, et al. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PLoS One 2009;

45. Cole ST, Brosch R, Parkhill J, et al. Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. Nature 1998;

46. Public Health England. England world leaders in the use of whole genome sequencing to diagnose TB - GOV.UK. 2017;

47. Fishbein S, van Wyk N, Warren RM, et al. Phylogeny to function: PE/PPE protein evolution and impact on Mycobacterium tuberculosis pathogenicity. Mol. Microbiol. 2015;

48. Akhter Y, Ehebauer MT, Mukhopadhyay S, et al. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: Perhaps more? Biochimie 2012;

49. Nanoporetech. Accuracy. 2020;

50. Illumina. Sequencing Quality Scores. 2020;

51. McIvor A, Koornhof H, Kana BD. Relapse, re-infection and mixed infections in tuberculosis disease. Pathog. Dis. 2017; 75:

52. Zagordi O, Bhattacharya A, Eriksson N, et al. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics 2011; 12:119

53. Anyansi C, Keo A, Walker BJ, et al. QuantTB – a method to classify mixed Mycobacterium tuberculosis infections within whole genome sequencing data. BMC Genomics 2020; 21:80

54. Wlodarska M, Johnston JC, Gardy JL, et al. A microbiological revolution meets an ancient disease: Improving the management of tuberculosis with genomics. Clin. Microbiol. Rev. 2015;

55. Schürch AC, Kremer K, Daviena O, et al. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J. Clin. Microbiol. 2010;

56. Schürch AC, Kremer K, Kiers A, et al. The tempo and mode of molecular evolution of Mycobacterium tuberculosis at patient-to-patient scale. Infect. Genet. Evol. 2010;

57. Gardy JL, Johnston JC, Sui SJH, et al. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. N. Engl. J. Med. 2011;

58. Estée Török M, Reuter S, Bryant J, et al. Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. J. Clin. Microbiol. 2013;

59. Stucki D, Ballif M, Bodmer T, et al. Tracking a tuberculosis outbreak over 21 years: Strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. J. Infect. Dis. 2015;

60. Miotto P, Tessema B, Tagliani E, et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. Eur. Respir. J. 2017;

61. World Health Organization. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance. 2021;

62. Liu Q, Zuo T, Xu P, et al. Have compensatory mutations facilitated the current epidemic of multidrug-resistant tuberculosis? Emerg. Microbes Infect. 2018;

63. Farhat MR, Shapiro BJ, Kieser KJ, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat. Genet. 2013;

64. Safi H, Lingaraju S, Amin A, et al. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-Arabinose biosynthetic and utilization pathway genes. Nat. Genet. 2013;

65. Sun G, Luo T, Yang C, et al. Dynamic population changes in mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients. J. Infect. Dis. 2012;

66. Cohen KA, Abeel T, Manson McGuire A, et al. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. PLOS Med. 2015; 12:e1001880

67. Feuerriegel S, Schleusener V, Beckert P, et al. PhyResSE: A web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. J. Clin. Microbiol. 2015;

68. Coll F, McNerney R, Preston MD, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;

69. Bradley P, Gordon NC, Walker TM, et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nat. Commun. 2015; 6:10063

70. Steiner A, Stucki D, Coscolla M, et al. KvarQ: Targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics 2014;

71. Coll F, Preston M, Guerra-Assunção JA, et al. PolyTB: A genomic variation map for Mycobacterium tuberculosis. Tuberculosis 2014;

72. Starks AM, Aviles E, Cirillo DM, et al. Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform. Clin. Infect. Dis. 2015;

73. Sandgren A, Strong M, Muthukrishnan P, et al. Tuberculosis drug resistance mutation database. PLoS Med. 2009;

74. Dai E, Zhang H, Zhou X, et al. MycoResistance: a curated resource of drug resistance molecules in Mycobacteria. Database 2019; 2019:74

75. Zignol M, Cabibbe AM, Dean AS, et al. Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. Lancet Infect. Dis. 2018;

76. Phelan J, O'Sullivan DM, Machado D, et al. The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. Genome Med. 2016;

77. Pankhurst LJ, del Ojo Elias C, Votintseva AA, et al. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: A prospective study. Lancet Respir. Med. 2016; 4:49–58

78. Metcalfe JZ, Streicher E, Theron G, et al. Cryptic microheteroresistance explains mycobacterium tuberculosis phenotypic resistance. Am. J. Respir. Crit. Care Med. 2017;

79. Metcalfe JZ, Streicher E, Theron G, et al. Mycobacterium tuberculosis subculture results in loss of potentially clinically relevant heteroresistance. Antimicrob. Agents Chemother. 2017;

80. Trauner A, Liu Q, Via LE, et al. The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy. Genome Biol. 2017;

81. Eldholm V, Norheim G, von der Lippe B, et al. Evolution of extensively drug-resistant Mycobacterium tuberculosis from a susceptible ancestor in a single patient. Genome Biol. 2014;

82. Koch A, Cox H, Mizrahi V. Drug-resistant tuberculosis: challenges and opportunities for diagnosis and treatment. Curr. Opin. Pharmacol. 2018;

2 QuantTB – A method to classify mixed *Mycobacterium tuberculosis* infections within whole genome sequencing data

2.1 Abstract

Background

Mixed infections of *Mycobacterium tuberculosis* and antibiotic heteroresistance continue to complicate tuberculosis (TB) diagnosis and treatment. Detection of mixed infections has been limited to molecular genotyping techniques, which lack the sensitivity and resolution to accurately estimate the multiplicity of TB infections. In contrast, whole genome sequencing offers sensitive views of the genetic differences between strains of *M. tuberculosis* within a sample. Although metagenomic tools exist to classify strains in a metagenomic sample, most tools have been developed for more divergent species, and therefore cannot provide the sensitivity required to disentangle strains within closely related bacterial species such as *M. tuberculosis*.

Here we present QuantTB, a method to identify and quantify individual *M. tuberculosis* strains in whole genome sequencing data. QuantTB uses SNP markers to determine the combination of strains that best explain the allelic variation observed in a sample. QuantTB outputs a list of identified strains, their corresponding relative abundances, and a list of drugs for which resistance-conferring mutations (or heteroresistance) have been predicted within the sample.

Results

We show that QuantTB has a high degree of resolution and is capable of differentiating communities differing by less than 25 SNPs and identifying strains down to 1× coverage. Using simulated data, we found QuantTB outperformed other metagenomic strain identification tools at detecting strains and quantifying strain multiplicity. In a real-world scenario, using a dataset of 50 paired clinical isolates from a study of patients with either reinfections or relapses, we found that QuantTB could detect mixed infections and reinfections at rates concordant with a manually curated approach.

Conclusion

QuantTB can determine infection multiplicity, identify hetero-resistance patterns, enable differentiation between relapse and re-infection, and clarify transmission events across seemingly unrelated patients – even in low-coverage (1×) samples. QuantTB outperforms existing tools and promises to serve as a valuable resource for both clinicians and researchers working with clinical TB samples.

2.2 Background

Tuberculosis (TB) - one of the oldest diseases in the world - continues to devastate the lives of millions per year. The World Health Organization's *End TB Strategy* calls for a 95% reduction of TB deaths by 2035, a feat that will require more innovative and effective methods to treat, control and diagnose the disease [1].

For centuries it was assumed TB patients were infected with a single strain of *Mycobacterium tuberculosis,* the causative bacteria of TB. However, molecular genotyping methods have illuminated the phenomena of *mixed infections* - sometimes also referred to as superinfections or co-infections [2–6]. Patients with mixed infections harbor multiple genetically distinct strains of TB at the same time. Previous research has suggested that mixed TB infections account for up to 30% of cases [4]. However, the real incidence largely remains unknown [7], with estimates ranging from 19% for sputum samples up to 51% for combinations of pulmonary and extra-pulmonary samples [5]. Mixed infections can complicate treatment and diagnosis through heteroresistance (presence of both drug susceptible and resistant patterns), which can cause false negatives in drug susceptibility tests and enable the spread of antibiotic resistance when left undetected [8–10]. Therefore, accurate detection of strains within a mixed infection, as well as their distinct resistance patterns, is important for decreasing the worldwide TB burden and slowing the spread of drug resistance.

Various molecular typing methods that can differentiate across the 8 major TB lineages, have been used to gain clues as to whether a particular infection contains more than one *M. tuberculosis* strain. Restriction Fragment Length Polymorphism (RFLP) analysis relies on the positioning and copy number of the variable transposable insertion element IS6110 [11]. Mycobacterial Interspersed Repetitive Unit-Variable Number Tandem Repeat (MIRU-VNTR) typing analyzes PCR amplified loci which vary in size and number of repeats [12]. Finally, spoligotyping analyzes a series of 43 spacer oligonucleotides in the directed repeat region [12]. As these methods only indicate the lineage(s) of the strain within a sample, they cannot identify intra-lineage infections, making them unsuitable for mixed infection classification. In addition, these approaches only examine a small portion of the genome, and were not originally intended for the detection of mixed infections.

In contrast, whole genome sequencing (WGS) offers a more comprehensive view into the genetic composition of a sample that includes distinct genetic information from individual strains. However, interpreting and analyzing such genomic data to identify and disentangle the composition of a mixed infection still remains a difficult task. To the best of our knowledge, few established methods exist to identify mixed infections for *M. tuberculosis* using WGS data. Some studies have classified a sample as mixed if the number of heterozygous positions (positions with evidence for more than one allele), exceeds a predefined arbitrary threshold [13,14]. These methods, which only consider mixes of two strains (bi-allelic variation), require sufficient coverage (>5x) for each allele and cannot be used to pinpoint actual strain identities. More recently, a paper by Sobkowiak et al [15], presents two methods, one based on the counts of heterozygous alleles and another based on a Bayesian framework to delineate strains. Neither method provides information on the identity of the strains, limiting their utility in comparing across samples, a valuable resource in transmission

studies or when differentiating relapse from reinfection. On the other hand, a previous method by Gan et al [16] classifies using a reference database. However their method and database is custom built for their own specific need and has not been made available or benchmarked. Other metagenomic tools exist to classify mixed populations of strains within a single species, such as Sigma, StrainEst, Strain Seeker, and Pathoscope [17–20]; however these tools were developed and benchmarked using bacteria with greater intra-species diversity, such as *Escherichia coli,* where high numbers of variable sites and strain-specific structural variations can be exploited to delineate strains. These methods were not designed to be able to discriminate between strains of highly clonal species like *M. tuberculosis*, where there is near perfect syntenic gene conservation, and typically much less than 2000 genome wide SNPs between the most genetically distant isolates, resulting in an average sequence similarity over 99.97% between any two independent isolates.

We present QuantTB, a tool that is specifically designed to identify and quantify the abundance of closely related *M. tuberculosis* strains in WGS samples containing TB at a detectable level, whether sourced from culture or sputum. QuantTB is highly relevant not only for TB research but also for diagnosis of TB in WGS data. Qualitative detection of mixed infections offers many benefits such as: characterizing hard to treat TB cases [21], facilitating analysis of seemingly unrelated transmission events involving lesser abundant strains, differentiating patients who have relapsed apart from those who harbor novel infections, and elucidating cases of poor treatment outcomes due to heteroresistance. In addition, QuantTB can readily be used in a diagnostic context, reducing processing time for TB identification in direct from sputum patient samples.

QuantTB classifies by iteratively comparing SNPs from an uncharacterized TB sample with a database of TB SNP profiles from known reference strains, resulting in a low rate of false positives, while retaining sensitivity at coverages as little as 1×. Unlike other tools that were designed for use on species with higher levels of intra-species variation, QuantTB can accurately and precisely disentangle TB strains that differ by as few as 25 SNPs. QuantTB also informs the user of any drug resistant or hetero-resistant loci within the sample.

QuantTB is available on GitHub: https://github.com/AbeelLab/quanttb/

2.3 Methods

2.3.1 Construction of a SNP-based reference database

QuantTB uses a reference database of SNP sequences for strain classification which is constructed in four steps: 1) selecting a broad set of TB genomes, 2) selecting representative SNPs within these reference genomes 3) filtering genomes based on SNP similarity, 4) addressing reference genome bias.

1. Acquiring genomes for the reference database

Although QuantTB can use either assemblies or raw sequencing reads for the construction of the reference database, assemblies are the preferred input. Assemblies represent aggregate, error-corrected versions of the corresponding read set and will yield superior results. We downloaded all available *M. tuberculosis* assemblies (5,867 complete and draft genomes as of July 23 2018)

from NCBI [22,23] using the taxonomic id: txid77643. We assigned lineages to each assembly based on lineage specific markers using a method described previously [24]. We filtered out 217 assemblies that did not associate with any known *M. tuberculosis* lineage. We removed 12 assemblies containing markers from more than one lineage, then confirmed the remaining genomes were of appropriate size, within a range of 4.4±0.5 MB. In total, 5,637 assemblies passed quality filtering. Supplementary Table 1 contains the NCBI accession codes and lineage prediction for all assemblies.

2. Selecting representative SNPs

Selecting high quality SNPs for each genome present in the reference database is paramount to the success of our method. QuantTB can extract SNPs from two different sources: assemblies (FASTA files or SNP files outputted by MUMmer's show-snps program (version 3) [25]) and read sets (FASTQ files or VCF files outputted by Pilon (version 1.22) [26]).

When extracting SNPs from assemblies, QuantTB aligns each assembly against the H37Rv reference genome (Genbank: CP003248.2) using MUMmer's nucmer command with the minimum cluster length set to 100 [25] and other parameters set to the default values. All outputted SNPs are used, except for those marked as ambiguous by MUMmer. In the analysis presented here, we extracted SNPs from the 5,637 reference assemblies that passed quality filtering for our reference database.

Although not used for the analysis presented in this manuscript, QuantTB can also extract SNPs from read sets. QuantTB aligns each read set against the H37Rv (Genbank: CP003248.2) genome with BWA-MEM (Version: 0.7.17-r1188) [27] using default settings, then index-sorts with samtools (Version: 1.6 , using htslib 1.6) [28]. By default, QuantTB uses Pilon (version 1.22, default settings with fixes set to none) [26] to generate a pileup and characterize each site. Sites denoted by Pilon as deletions, insertions, low coverage, and reference calls are excluded, in addition to low quality sites (Phred quality score less than 11), and ambiguous sites (alternate allele frequencies less than 0.9).

For SNPs from both assemblies and read sets, we applied a number of additional filters. SNPs within a specified distance from one another (default 25bp) were removed from consideration, as these could be indicative of sequencing or alignment error. QuantTB also excludes all variants that are located in genes annotated as PE/PPE (Supplementary Table 2) within the H37Rv reference, as these genes are known to be highly repetitive and prone to mapping errors, making it difficult to call variants using short-read data [29–31]. The resulting SNP sequence for a genome is a dictionary of positions *(p)* that differ from the H37Rv genome mapped to their corresponding alleles, where $allele(p_x) \rightarrow \{A, C, G, T\}$. The complete collection of SNP sequences in the reference database is stored in a binary matrix, where rows are the genomes and columns are the locus/allele pair (Figure 2.1).

3. Filtering genomes based on sequence similarity

The last step in constructing the reference database is to remove highly similar genomes. We calculated the pairwise SNP distances between each genome pair by summing the number of SNPs unique to each genome, i.e. by taking the union of variants minus the intersection of variants. If the SNP distance was below a specified threshold, the genome with the lowest number of SNPs

was removed. This process was repeated until all genomes differed by the specified minimum SNP distance. We evaluated the performance of QuantTB by constructing reference databases with four different SNP distance thresholds: 10, 25, 50 and 100 SNPs. Table 2.1 shows the number of strains within each reference database.

Name	Minimum Genomic Distance (SNPs)	Number of genomes
d10	10	4,933
d10small*	10*	200*
d25	25	3,686
d50	50	2,843
d100	100	2,167

Table 2.1 The number of genomes in each database after filtering by SNP distance. The distance was calculated by summing the number of unique SNPs between genomes. *In order to have a smaller database to benchmark against slower/more memory intensive tools, the number of genomes in d10small was restricted to be 200. The 200 genomes were randomly selected relative to the overall distribution of lineages, with a minimum requirement of five genomes for each lineage. D10 was selected as source set for the small benchmarking set to ensure the broadest possible strain and distance representation.

4. *Addressing reference genome bias*

All SNPs were called using the reference genome, H37Rv, introducing a bias that strains highly similar to the reference genome become 'invisible' using this method, because they have a very low number of SNPs. To remedy this issue, a custom SNP-based representation of the H37Rv sequence was generated, based on the frequencies of SNPs across all other genomes in our reference database. If the same variant is observed in almost all the genomes in the reference database, we designate this as an H37Rv specific variant, i.e. a SNP within the H37Rv genome compared to every other genome. Therefore, QuantTB generates an "H37Rv SNP sequence" including positions where more than 75% of the genomes in the reference database have a common allele that differs from H37Rv. These locations are a fingerprint for H37Rv-like strains to identify them from the rest of the database.

2.3.2 Using the SNP database to quantify strains present within a sample

QuantTB uses a SNP-based reference database to process short-read data in order to quantify the set of strain(s) present within a sample, such as short-read data from a clinical sample or isolate. Sample processing is done in two steps: 1) Extracting SNPs from a sample 2) Iterative classification of strains in the sample.

1. *Extracting SNPs from a sample*

QuantTB can accept either a FASTQ file or a VCF file as an input sample for classification. Given a FASTQ file, reads are aligned against the H37Rv genome using BWA-MEM with default settings.

A pileup is generated using Pilon with the default parameters and fixes set to none. Insertions, deletions, bases with low quality (Phred less than 11) and bases within PE/PPE regions are removed as in the construction in the reference database. All other bases with a frequency greater than 0.99 for the reference allele are removed. The end result is a dictionary containing the extracted allele coverages and frequencies for every SNP position identified in the database. Note that QuantTB does not filter based on coverage; this allows for the detection of low abundance strains within a sample.

2. *Iterative classification of strains in the sample*

Specific TB strains within the reference database are identified as present within a sample by iteratively querying against the SNP-based reference database. Figure 2.1 shows an example of this iterative process in a mixed sample. The steps of the algorithm are as follows:

I. Compute a "strain presence score" (s_i) for every genome *(i)* in the database (see below for computation of score).
II. Choose the genome with the highest strain presence score, s_i.
III. Remove the chosen genome's SNPs from the database and sample.
IV. Repeat steps 1-3 until no more SNPs remain, the strain presence score is below the threshold, or the maximum number of iterations have been reached.

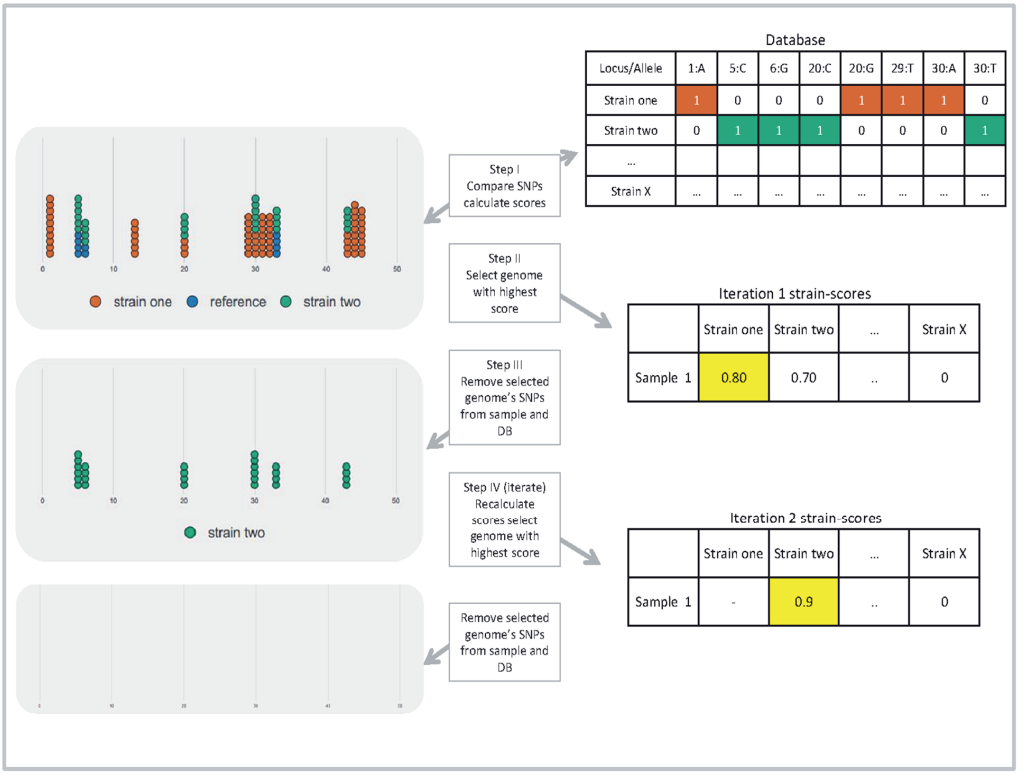


Figure 2.1 Iterative multiple strain identification process in QuantTB for a mixed sample, where two strains are present , strain 1(red) and strain 2 (green). First, SNPs from the sample are compared against SNP sequences in the reference database to calculate a strain presence score for every genome in the database. The sample is represented as a pileup, where every circle represents an allele copy. Red circles indicate alleles unique to strain A, green indicates alleles unique to strain B, and blue indicates reference strain (blue). The database (top right) is an example matrix representation of a reference genome database. Each column represents a single SNP (unique position and variant), and each row represents a genome in the reference database with this SNP present (1) or absent (0). Strain presence scores are calculated for every genome in the reference database. The genome with the highest strain presence score (s_i) is selected, in this case strain A (red). The SNPs associated with strain A are removed from the database and the input sample, along with additional reference alleles. In each subsequent iteration the scores are recalculated, allowing for the identification of additional strains, and the process continues until there are no more SNPs or a threshold has been reached.

Computation of strain presence score.

During each iteration, a strain presence score *(s_i)* is calculated for every genome in the database (*D).* The strain presence score is an average of two statistics, O_i and A_i, and represents the overall presence of a strain within the sample. O_i and A_i are described below.

$\boldsymbol{O_i}$ represents the fraction of SNPs from a particular reference genome, *i*, that was observed in the sample. The higher O_i, the more likely the set of SNPs observed in the sample originated from genome *i*.

$$O_i = \frac{|Al_{sample} \cap Snps_i|}{|Snps_i|}$$

Al_{sample} is the set of alleles observed above a coverage threshold t_a. Applying a coverage threshold diminishes the effect of random errors in the sample, while retaining sensitivity for true variation. This threshold t_a, is dynamic and determined by the average coverage of the sample, C_{sample}, and the average coverage of the genome identified in the previous iteration, $C_{G_{k-1}}$.

$$t_a = \begin{cases} \max(2, 0.05 \times C_{G_{k-1}}) & \text{if } C_{sample} > 25 \\ 0.05 \times C_{G_{k-1}} & \text{if } C_{sample} \leq 25 \end{cases}$$

If the sample has an average coverage greater than 25, a minimum coverage threshold of 2 is set for all iterations, whereas for samples with an average coverage less than 25, there is no minimum, so that strains at low coverage can still be detected. For each iteration *k*, the threshold is set as 5% of the average coverage of the strain identified in the previous iteration. This is initialized at *k=0* as 5% of the sample coverage (C_{sample}). Applying a coverage threshold diminishes the effect of random errors in the sample, while retaining sensitivity for true variation. Notice that this threshold likely goes down in every iteration as the coverage of the previously detected strain is used with a minimum of 2.

$\boldsymbol{A_i}$ represents the frequency with which a particular genome's SNPs accounts for all the allelic variants present in the sample. The previous statistic, O_i, represents how many SNPs of a particular genome been observed with sufficiently high coverage. However, when a sample has low coverage, the probability of observing the complete set of a genome's SNPs is low. To account for strains present at low coverages, QuantTB also calculates, A_i.

$$A_i = \frac{|Freq_i|}{|Al_{sample}|}$$

Where $Freq_i$ represents the vector of frequencies for each allele of genome i within the sample:

$$Freq_i = \left(f_{p_{i,1}}, f_{p_{i,2}}, f_{p_{i,3}}, \ldots, f_{p_{i,L}}\right), \quad f_x \in [0,1].$$

Choose the genome with the highest strain presence score

At the end of each iteration, the strain presence score $(s_i,)$, is calculated as an average between O_i and A_i, and the genome with the highest s_i,is selected as being present in the sample.

Remove the chosen genome's SNPs from the database and sample

Before the next iteration begins, SNPs corresponding to the chosen genome are 1) removed from each SNP sequence in the database and 2) removed from the sample. In addition, any H37Rv alleles present in the sample at positions outside of the identified genomes' SNP sequences are also removed. This is because those alleles can be accounted for already by the presence of the identified genome.

Because it is unlikely that the true strain present in the sample shares the exact collection of SNPs with its highest scoring match in the database, additional SNPs from the sample could match erroneously across multiple other genomes in the database with enough coverage to be marked as 'observed'. As the coverage increases, the probability that an additional genome is spuriously detected also increases, due to the number of these uninformative SNPs that do not match perfectly with the originally selected genome. QuantTB implements a check to safeguard against this. To account for spuriously detected genomes due to higher coverages (greater than 25), we only allow strains to be detected in a sample when their prevalence accounts for at least 1% of the sample coverage. Therefore, SNPs from a particular strain are only removed from the sample when the change of coverage at each iteration would be at least 1%, otherwise the strain is ruled out for detection.

Iteration

The QuantTB algorithm iterates until the score threshold has been reached (the default is 0.15 but this can be adjusted by the user). Before starting the next iteration, a check is performed to ensure that a sufficient number of SNPs (15) still remain in the sample and in the database for reliable classification. This value was empirically determined during large scale testing.

At the end of the iterations, relative abundance is calculated by taking the average coverage of unique SNPs for each genome in the sample.

2.3.3 Prediction of antibiotic resistance status of detected strains

In order to identify presence or absence of a resistance phenotype in the sample, QuantTB uses a curated set of SNPs conferring antibiotic resistance to seven TB drugs generated from the previous study of Manson et al [24] (Supplementary Table 3). QuantTB also allows users to upload their own curated set of SNPs. If resistance conferring allele(s) are present at a frequency of more than 90%, the sample is considered fully resistant for that drug. Heteroresistance, where there is evidence of both a resistant and a susceptible phenotype in a sample, can occur due to mixed infections or through in-host microevolution. If a resistance conferring allele(s) is present at a

frequency between 10-90%, then the sample is considered heteroresistant for that drug. QuantTB outputs the results of the resistance testing in a separate file, if the appropriate command-line flag is set.

2.3.4 Benchmarking using synthetic read sets

We constructed test datasets to benchmark QuantTB and compare its performance to two other strain level identification methods, StrainSeeker [18] and Sigma [17]. Another tool, StrainEst [32] is also capable of performing single strain classification; however, a downloadable script is not provided to construct a database for *M. tuberculosis* genomes compatible with their algorithm, so we were unable to include it in our benchmark.

Synthetic mixed samples of two and four strains were used to perform benchmarking. In order to benchmark overall performance across different coverage levels, as well as across databases with different levels of strain similarity, we constructed mixes of four strains, where all four strains were present at equal relative abundance. In order to further benchmark the ability of QuantTB to assess samples containing strains with different relative abundances, we generated synthetic mixes of two strains sampled at different relative abundances.

To generate the four strain mixtures we randomly selected 200 combinations of four assemblies from each of the four reference databases generated with different SNP-distances using publicly available *M. tuberculosis* assemblies. In total, we selected 800 different combinations of four strains. For each reference database, we ensured that all 7 main lineages were represented across the selected sets of assemblies. Then, for each selected assembly, we synthesized paired end reads using ART (Version 2.5.8) [33] with default settings for the Illumina HiSeq 2500 platform, at a read length of 101 bp and a final coverage of 100×. Each read set was down sampled to 0.1×, 1×, 10×, and 20× coverage, then merged into mixes of four. This corresponds to 800 mixed sets of four different coverage levels, or 3200 synthetic mixes of strains.

To generate synthetic two-strain mixtures of strains at different relative abundances, we randomly selected 100 pairs of assemblies from each of the d50 and d100 reference databases. Paired end reads were simulated for each assembly, then the read sets were merged in mixes at 1×/9× coverage and 3×/7× coverage. This corresponds to 200 mixed sets at two different coverage levels, resulting in 400 synthetic mixes of varying relative abundance.

In addition, we generated synthetic four-strain mixtures for a smaller dataset, able to run in shorter compute time. StrainSeeker and Sigma are not capable of processing large sized reference sets (>2000 genomes) and required >3 days of compute time per sample or >7 days for reference database construction of 2,000 genomes. Therefore, to compare the performance of QuantTB against that of StrainSeeker and Sigma within a reasonable time frame, we created a smaller reference database, d10small. Using the reference genomes from the d10 database (see Methods), we randomly selected 200 genomes such that each TB lineage was represented in proportion to its relative incidence in the overall dataset, with a minimum requirement of five representatives for each lineage. Synthetic sample sets were then created based on the small reference set, using 200 randomly selected sets of 4 genomes. These sets were synthesized using the same method as for the previous databases, with the only exception being that we only created samples where the strains are present at either 1× and 10× coverage.

2.3.5 Benchmark evaluation using synthetic sets

In order to test the performance of each method, we calculated the *Recall*, *Precision*, and the *F1 score* for every test category. True positive (*TP*) refers to the number of correctly identified strains. False positive (*FP*) refers to the number of identified strains that were not present in the sample. False negative (*FN*) refers to the number of strains present in the sample that were not identified.

$$Recall = \frac{TP}{TP + FN}, \qquad Precision = \frac{TP}{TP + FP}, \qquad F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

2.3.6 Evaluation using real genomic data

We demonstrated the utility of QuantTB with real data samples from a study investigating reinfection and relapse using WGS [13]. Sequencing reads from 50 pairs of isolates were downloaded from the SRA [34]. SRA files were extracted using fastqdump (Version 2.9.0) [34] from the SRA toolkit, using the "split-3", "skip-technical", and "clip" flags to split left and right reads into separate files, remove technical reads, and clip off poor-quality ends of reads, respectively.

To construct a phylogenetic tree from these samples, SNPs were extracted and filtered as described above. FastTree [35] was used to generate a tree from the concatenated SNPs.

2.4 Results

2.4.1 Comprehensive TB reference database captures the breadth of the *Mycobacterium tuberculosis* species

QuantTB requires a reference database of known *M. tuberculosis* genomes for classification, where every genome is represented by a set of SNPs (see right panel in Figure 2.1). To construct a TB reference database, we used 5,637 assemblies from NCBI which passed our quality filters (see Methods).

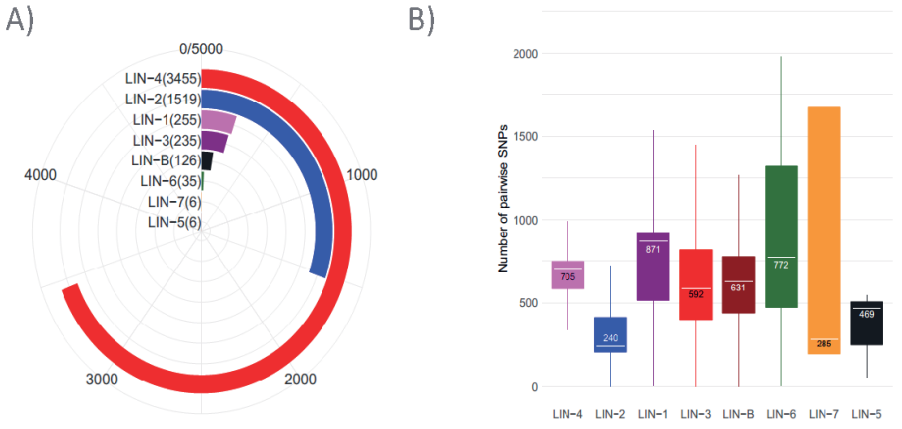


Figure 2.2 Panel A) Number of representatives from each lineage amongst all 5,637 *M. tuberculosis* assemblies in our reference database. Panel B) Intra-lineage pairwise distance for each lineage as measured by the number of unique SNPs between a pair. The number in the box plot is the median distance of all pairs of samples from that lineage.

Our database contained eight major lineages of TB at frequencies reflecting the overall abundances of sequences for each lineage in NCBI (Figure 2.2). Lineage 4 strains encompass the vast majority of *M. tuberculosis* assemblies currently available at NCBI (3,455 strains), while lineage 7 and lineage 5 are the least abundant with 6 strains for each (Figure 2.2A). The genetic diversity within lineages (Figure 2.2b) was in agreement with previous studies (33): (i) lineage 1 had the greatest intra-lineage genetic diversity (median of 871 SNPs pairwise distance) and (ii) lineage 2, the second most frequently occurring lineage, had the lowest diversity, (median of 240 SNPs pairwise distance). The six strains that comprise lineage 7 had a wide range of genetic diversity, suggesting the need for increased sequencing of less well-characterized lineages, which would improve the resolution of classification within these less abundant lineages.

To benchmark QuantTB's performance across databases with varying intra-database genetic distances, we constructed a set of databases with differing minimum differences between strains (10, 25, 50 and 100 SNPS, Table 2.1, Methods). Each database contained a representative distribution of strains from each lineage (Supplementary Table 4), as well as representative genetic diversity within each lineage (Supplementary Figure 1, Figure 2.2b) [36]. There was good concordance between the diversity represented in the complete data set (Figure 2.2b) and the derived benchmarking sets (Supplementary Figure 1).

2.4.2 QuantTB outperforms other tools using simulated data

We compared QuantTB's ability to accurately identify strains with Sigma [17] and Strainseeker [18]. We used five reference databases that varied both in size and in the genetic distance between representative genomes (Table 2.1). As Sigma and StrainSeeker are more computationally expensive than QuantTB, we were not able to use our larger databases of mixtures of four strains

(>200 strains) with tools other than QuantTB. In contrast, QuantTB scaled well with database size: database construction was complete in less than two hours, and a sample took less than 20 minutes on average to process using the same computer hardware. The ability to take advantage of a large reference database is a substantial advantage for QuantTB over StrainSeeker and Sigma, since the number of publicly available TB sequences in NCBI that could be included in the database is increasing rapidly. In future, an even larger database could allow for even finer resolution strain detection.

We compared performance of QuantTB with StrainSeeker and Sigma, using a smaller, lower-resolution database of 200 strains (d10small). While StrainSeeker performed on par with QuantTB (Figure 2.3A), both achieving near perfect F1 scores at both coverage levels, Sigma did not perform as well. Sigma identified the correct strains in almost all cases; however, this was accompanied with greatly reduced precision (Supplementary Table 5), i.e. including many false positives and decreasing its overall F1 score (Figure 2.3A).

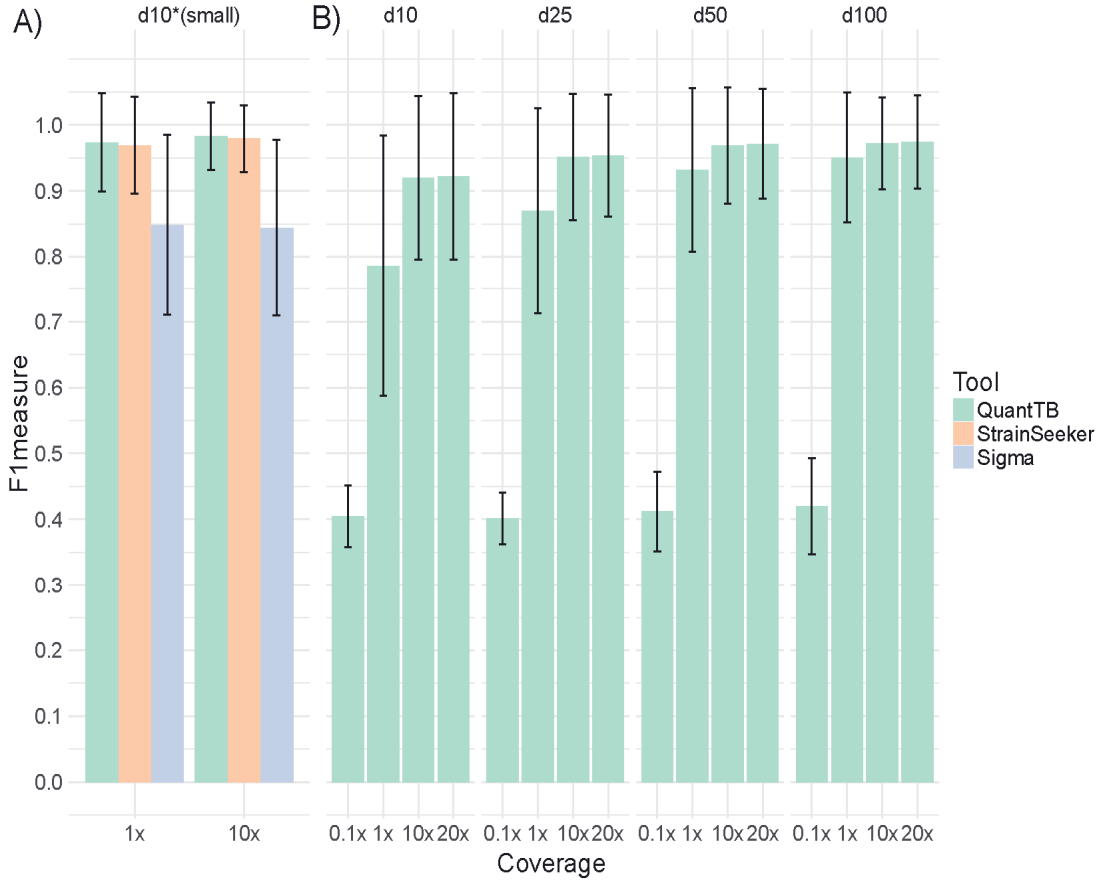


Figure 2.3 Benchmarking results of synthetically mixed read sets of three different strain identification tools, QuantTB, StrainSeeker and Sigma. A) Results from a smaller database (d10small, n = 200) are shown for

all tools for coverage levels of 1× and 10×, B) results from four larger databases (see Table 2.1) are shown only for QuantTB, for coverages ranging from 0.1× - 20x.

Classification of synthetic four-sample mixes using the larger reference databases presented a more difficult task; however, QuantTB's performance remained high (Figure 2.3B), achieving F1 scores above 0.9 at all coverages above 1x per strain, indicating that QuantTB was almost always able to predict all four strains in the synthetic mixes correctly. Scores for lower sample coverage (0.1× per strain) were reduced (F1 score of 0.4). The decreased SNP counts in these very low-coverage simulations led QuantTB to predict only one of the strains present for most of these samples (Supplementary Table 5) We also observed that samples of 20× coverage per strain performed just as well as samples of 10× coverage per strain, indicating no gain in performance from additional coverage. At 1× coverage per strain, QuantTB still performed adequately, with only a slight performance dip noticeable in the largest database containing 4,933 strains differing by at least 10 SNPs. We observed that the lower performance occurred mostly because QuantTB would predict a genetically similar strain instead of the *correct* strain. Taken together, these results suggest that QuantTB can detect strains present at a minimum of 1× coverage. In addition, the fact that the QuantTB algorithm efficiently scales to larger databases not only shows it can accurately classify genomes regardless of database content, but that it runs sufficiently fast to provide the required quick turnaround time in a clinical setting using a large, clinically representative database.

2.4.3 QuantTB accurately predicts relative abundances

To assess the ability of QuantTB, StrainSeeker, and Sigma to correctly predict relative strain abundances, we simulated mixed samples of pairs of strains that varied in their relative proportions (Figure 2.4). The setup represented a more realistic scenario, where strains in the samples (sourced from the d50 database) were not already present in the database (d10small). Additionally, we tested QuantTB's ability to predict relative abundance across different datasets of known strain identity (Figure 2.4A). Data for StrainSeeker and Sigma were not shown for this experiment due to inability to construct the larger reference databases.

QuantTB was by far the most successful tool at identifying the correct number of strains. QuantTB identified the correct number of strains (two) in the majority of samples (72%). StrainSeeker usually underestimated the number of strains and was only able to identify the correct number of strains in 25% of cases. Sigma failed to predict the correct number of strains in any sample, predicting at least 9 strains for all of the samples ((Figure 2.4B). For samples where QuantTB correctly predicted the strain multiplicity, it also predicted relative abundances close to the expected values, performing best for samples with a 0.1/0.9 strain ratio ((Figure 2.4, left graphs).

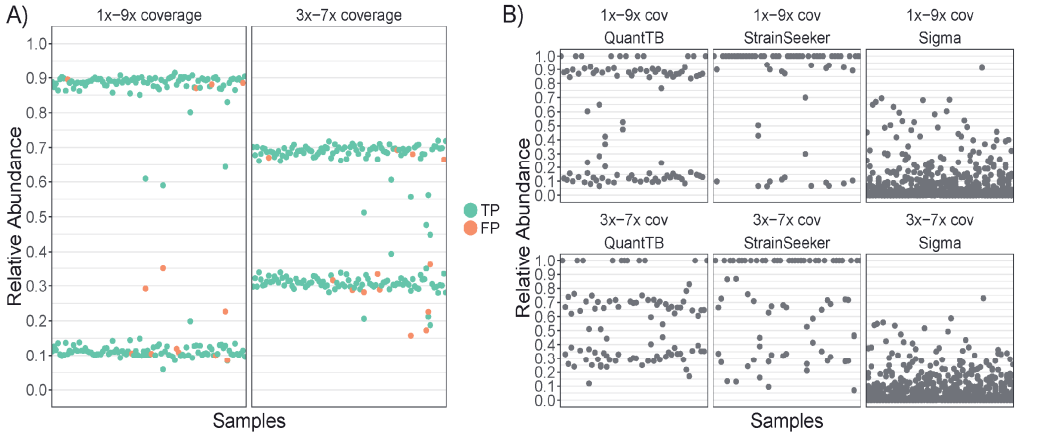


Figure 2.4 A) Relative abundance predictions across the synthetic sample sets, using randomly selected strains from the d50 and d100 database for QuantTB only. If the strain was correctly predicted for the sample it is colored green (true positive), whereas incorrectly predicted strains are colored red (false positive). The left graph contains samples where two strains are present at 1× and 9× coverage. The right graph contains samples where two strains are present at 3× and 7× coverage. B) Predicted relative abundances across synthetically mixed samples for QuantTB, StrainSeeker and Sigma. Each point represents a predicted relative abundance for a single strain. Each mixed sample contained a pair of strains from the d50 dataset at either 1×-9× or 3×-7× abundance. Although samples were sourced from the d50 dataset, the tools used a different set of genomes as a reference set (sourced from d10). Thus genomes in the samples were not present in the underlying database the tools were trained on. This lets us see how well each tool is at predicting the correct number of strains and the correct relative abundance between strains if the 'correct' strain in the sample is not already present in the database.

It is not only important to determine whether a tool is able to predict the correct abundances, but also whether it can select the most appropriate genome when the correct strain is absent from the reference database. Therefore, as genomes from the d50 database were used as test samples and tested against genomes in the d10small database, we evaluated the accuracy of strain predictions by assigning a true positive to each strain in a sample if QuantTB predicted the 'correct' relative genome in the d10small database (i.e. amongst the top 3 genomes with the highest pairwise SNP distance to the original strain). We found that QuantTB predicts the closest strain to the actual genome with an average precision value of 95%. This more realistic scenario, with previously unseen strains, suggests that QuantTB is able to accurately predict the correct number of strains even in cases where a near-identical strain is not already present in the database. Predictions of Sigma and StrainSeeker for strain multiplicity and relative abundances were insufficiently accurate (Figure 2.4B and Table 2.2) to perform this analysis meaningfully.

As only QuantTB could process samples using the larger databases, we further tested its accuracy at identifying correct strain pairs and their differing relative abundances using the d50 and d100 databases. For both databases, QuantTB accurately classified the identity of each strain in the pair (F1 measure of 0.98 and 0.92 for the d100 and d50 databases, respectively, Supplementary

Table 5) and accurately determined the relative abundance for each strain in the pair ((Figure 2.4B). The majority of relative abundances predicted were within 0.05 of the correct value (Supplementary Figure 2). Even in the few cases where QuantTB predicted the incorrect strain, QuantTB predicted it to be present in the sample at the correct relative abundance.

2.4.4 QuantTB differentiates between relapse, reinfection, and mixed infections in real world data

To demonstrate QuantTB's utility for (clinical) research, we quantified the distribution of *M. tuberculosis* strains within samples from a study investigating the frequency of TB relapses within patients from the REMoxTB clinical trial, a trial which evaluated treatment for TB in previously untreated patients [13]. Bryant et al. sequenced 50 pairs of isolates, one taken at an initial time point and the other taken after more than 17 weeks of treatment. Some samples were sequenced more than once (105 total sequencing datasets). Since there are no established methods for detection of mixed infections in *M. tuberculosis* genomic data, the original study used manual inspection of heterozygous SNPs to differentiate between relapse (same infecting strain), reinfection (a different infecting strain) and mixed infections. In the original study, a sample was labeled as mixed if the number of heterozygous loci exceeded a threshold, and as a reinfection if the SNP distance between pairs exceeded a threshold.

Here, we systematically reanalyzed this data using QuantTB and compared our findings from this dataset to those of Sigma and StrainSeeker. As it is impossible to know the identity of the strains present in the real samples in advance, we limited analysis to the multiplicity, or the number of strains identified in each sample. Table 2.2 shows the multiplicity of infection detected across the dataset of 105 samples for QuantTB, Sigma and StrainSeeker.

Number of predicted Strains	QuantTB (d10*)	StrainSeeker (d10*)	Sigma (d10*)	QuantTB (d10)	QuantTB (d25)	QuantTB (d50)	QuantTB (d100)
1	96	1	1	94	96	95	94
2	9	0	0	11	9	10	11
5 to25	0	6	21	0	0	0	0
26 to 45	0	90	65	0	0	0	0
46 to 60	0	2	8	0	0	0	0

Table 2.2 Number of samples predicted to contain the specified number of strains, using different methods and databases, for the set of 105 samples from Bryant et al.

QuantTB reported a consistently low (0-2) number of strains, and identified the same seven samples as mixed, irrespective of the database used as a reference, which was in agreement with the expected strain multiplicity based on Bryant et al. In contrast, StrainSeeker and Sigma reported an unrealistically large number of strains (greater than 25 on average).

By applying the results from QuantTB we were able to classify each sample as either part of a relapse, a reinfection or a mixed infection (4 cases). We used results from the d25 database because it performed optimally in our benchmarking tests. If more than one strain was identified by QuantTB, the sample was marked as a mixed infection. If the same strain was identified for both isolates in a pair, the sample pair was marked as a relapse case (35 cases). Finally, if different strains were identified across pairs, the sample pair was marked as a reinfection (3 cases). Bryant et al. removed 9 samples due to contamination but did not mention how how contamination was determined. Although cross contamination from other species would have had little impact on QuantTB (given sufficient TB coverage), we also removed these samples during our analysis.

Category	QuantTB	Bryant et al.
Mixed (Lost a strain at later time point)	Sample 2: 81%-18%	Sample 2
	Sample 8: 53%-47%	Sample 8
	Sample 23: 60%-40%	Sample 23
	Sample 50: 75%-25%	Sample 50
Mixed (Gained a strain at later time point)		**Sample 42**
		Sample 45
Reinfection	Sample 10	Sample 10
	Sample 14	Sample 14
	Sample 35	Sample 35
	Sample 15	
Relapse	33 matching samples **Sample 42** **Sample 45** **Sample 3**	33 matching samples
Clinically TB negative on follow up	Sample 36 (H37Rv) Sample 37 (H37Rv) Sample 38 (H37Rv)	Sample 36 Sample 37 Sample 38 **Sample 3** **Sample 15**

Table 2.3 Comparison of all mixed infections, reinfection and relapses called between QuantTB and Bryant et al. Samples in bold are discordant between the two methods. QuantTB predictions also include the abundance levels of both strains identified within the sample. Samples labeled as Clinically TB negative on

follow up were cases in which the second of the isolate pair assigned to the H37Rv strain by QuantTB, and tested negative for TB in the original study.

The manual analysis of Bryan et al. designated six samples as mixes. The results from QuantTB match those of Bryant et al for the vast majority of cases (Table 2.3), classifying the same 3 samples as reinfections, 4 samples as mixed infections, and 33 samples as relapses. QuantTB classified three additional samples as relapses. Samples 42 and 45 were identified as mixed infections in the original study. Upon investigation, it was found that the original study labeled these as mixed infections not based on their original threshold but based off of a 'manual inspection', which was not well described. Sample 3 was manually identified as a 'single isolated positive', a label given when the second isolate of a pair tested negative for *M. tuberculosis* under culture. Four additional samples were given this label by Bryant et al., who mentioned that these cases were mostly caused by cross contamination. In three culture negative samples labeled 'single isolate positive' by Bryant et al., QuantTB identified H37Rv (a laboratory strain). As the coverage for the H37Rv reference strain was high in these three samples, our analysis supports the hypothesis that three culture negative isolates resulting in the sequencing of the H37Rv laboratory strain. The remaining discrepancy, Sample 15, was classified as a reinfection by QuantTB instead of a single isolated positive.

To further validate our predictions and clarify discrepancies with the original study, we constructed a phylogenetic tree of all 105 sample isolate pairs based on concatenated SNP sequences (see Methods). This allowed us to visualize the phylogenetic distances between isolates of a sample pair (Figure 2.5). We observed that most sister leaves in the tree were part of the same sample isolate pair, representing relapse cases. The two samples classified as mixed by the original study but as relapses by QuantTB also appear as sister nodes on the tree ((Figure 2.5, boxes A.1 and A.2). Although this does not rule out a mixed infection, it justifies QuantTB's relapse classification. In addition, we observed the clustering of isolates which QuantTB identified as most similar to H37Rv (purple nodes in (Figure 2.5, box B), which were classified as 'single isolated positive' by the original study. The other samples given this designation by the original study, Sample 3 and Sample 15, did not have an isolate clustered with the H37Rv strain. Instead Sample 3's isolates were sister nodes on the tree ((Figure 2.5, box C) and the two isolates of Sample 15 were found on opposite ends of the tree ((Figure 2.5, boxes D.1 and D.2), both locations confirm QuantTB's predictions of relapse and reinfection, respectively.

Finally, we observed two samples whose isolate pairs appeared swapped on the tree: Sample 2 (mixed infection) and Sample 10 (reinfection). Sample 2A has sister nodes with Sample 10B (box E.1), while Sample 10A has sister nodes with Sample 2B on a distant part of the tree (box E.2). Before treatment, Sample 2 (isolate 2A) was mixed with two strains, the minor of which was present within isolate 10A. After treatment, the major strain of Sample 2 was lost, leaving the second pair of Sample 2 (isolate 2B) with only the minor strain, explaining its change of location (next to isolate 10A) on the tree. On the other hand, after treatment, the patient carrying sample 10 was re-infected with a different strain that was similar to the major strain of isolate 2A. Without the annotation of QuantTB it would appear a sample swap might have occurred. But with QuantTB

this occurrence can be explained by reviewing the strain identities, because QuantTB outputs which genome has been detected in the sample.

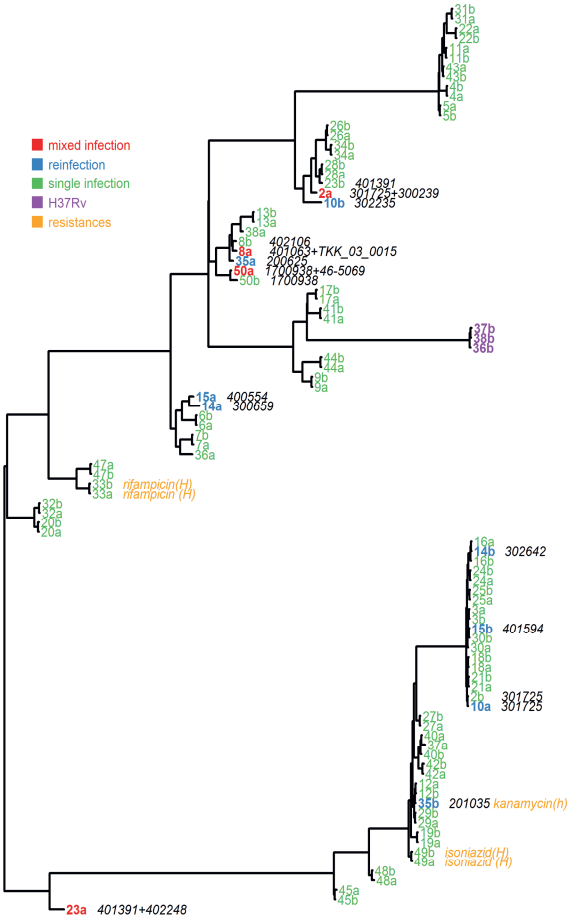


Figure 2.5 Phylogenetic tree of 47 pairs of isolates from sequencing reads taken from the study of Bryant et al. Tips are labeled with the isolate number and its part of the pair (a or b), and are colored by its isolate classification as predicted by QuantTB. Isolates containing a mixed infection are colored in red. Isolates part of a reinfection pair are colored in blue. Isolates containing the H37Rv strain are colored in purple. Isolates containing antibiotic heterozygous (h) or homozygous (H) resistance mutations are in orange. All single infections isolates are colored in green. To the right of the mixed and reinfection isolates, we show the strains present in the isolate as predicted by QuantTB. Boxes are discussed in the main text.

Overall, QuantTB and the manual curation presented in the original study resulted in agreement for 43 of the 47 sample predictions (91%). In the remaining cases, we have presented reasons why QuantTB's prediction may be at least as accurate as the original manual designations. In

addition, QuantTB gives information that was not available from the manual approach of Bryant et al., including detail on multiplicity of infection, and the identity and abundance of each strain, giving a detailed overview of each sample's genetic makeup.

2.5 QuantTB provides insight into antibiotic resistance

Using QuantTB, we determined the antibiotic resistance genotype for each of the isolates. Antibiotic resistance was indicated if the sample had a SNP in one of the antibiotic resistance causing loci from a previously published curated list (see Methods) [24]. Heteroresistance was indicated if the sample had alleles supporting both the resistant and susceptible genotype at a particular locus. Bryant et al. also tested for antibiotic resistance, both phenotypically (with mycobacterial growth indicator tube susceptibility testing) and genotypically (their method was not described). They found no evidence of genotypic or phenotypic antibiotic resistance in any sample. However, we found evidence for genotypic antibiotic resistance in five isolates (Table 2.4, (Figure 2.5). Two isolates were from the same patient, 33 and 49 (relapse cases) while one was the second isolate in its sample pair, 35b (reinfection case). We found no relation between mixed infections and heteroresistance, nor do we find evidence of the emergence of antibiotic resistance within a relapse case. Isolate 35b exhibited heteroresistance to kanamycin in one locus: 13% of alleles were of the resistance phenotype, and 87% were susceptible. Because this was a reinfection case, it is not possible to determine whether the heteroresistance arose due to within host evolution.

Isolate	Drug	Gene: Amino Acid position, Amino acid change	Type	distribution
33a	rifampicin	RpoB: His-445-Ser	homozygous resistant	
33b	rifampicin	RpoB: His-445-Ser	homozygous resistant	
35b	kanamycin	Intergenic: MurA-Ogt	heteroresistant	res: 0.13% sus: 0.87%
49a	isoniazid	KatG: Ser-315-Thr	homozygous resistant	
49b	isoniazid	KatG: Ser-315-Thr	homozygous resistant	

Table 2.4: Isolates exhibiting genotypic antibiotic resistance from the Bryant et al. dataset

2.6 Discussion

Mixed infections are known to complicate treatment and diagnosis of tuberculosis [8–10]; however, the true clinical impact and prevalence of mixed infections is still poorly understood due to the lack of suitable methods to detect and quantify individual strains of *M. tuberculosis*. WGS studies investigating *M. tuberculosis* typically identify mixed infections based on the amount of heterozygous base calls [6,13,14,37]. However, both the definition of a heterozygous locus and the number of heterozygous positions indicative of a mixed infection varies between studies. For example, Bryant et al. defined a position as heterozygous when two alleles were supported by at least 5% of the reads with a minimum read depth per allele of 4, and a sample as mixed if it had more than 80 heterozygous base calls [13]. Guerra-Assunção et al. defined a position as heterozygous if it had at least 30× coverage and more than one allele accounted for in at least 30% of the reads, and classified a sample as mixed if more than 140 bases were heterozygous [14]. Perez Lago et al. simply called a position heterozygous when the less frequent allele was supported by 5 reads [37]. With QuantTB we aimed to provide analgorithm capable of systematically quantifying the multiplicity and abundance of *M. tuberculosis* strains at high resolution using WGS data that does not require manual definitions or counting of heterozygous positions. Because of QuantTB's unique algorithm that identifies strains in an iterative process, strains can be detected at low coverages (1×), irrespective of the relative frequencies of alleles. The information provided by QuantTB provides several key improvements over a manual approach of counting heterozygous positions. QuantTB: 1) outputs the specific identity of the strain, making the tracking of specific strains across samples possible; 2) outputs the abundances of every strain identified in the sample, enabling the quick identification of major and minor subpopulations; 3) is capable of detecting more than two strains; 4) reduces time required to identify mixed infection to less than half an hour; 5) simplifies the manual process into an easy to use package reducing the amount of overhead; and 6) increases reproducibility across TB studies.

Due to QuantTB's use of a reference database, tracking the presence or absence of specific strains across a set of longitudinal or outbreak samples is also possible. Within a sample, QuantTB can identify the closest strain(s) present from a reference database, even using a large database containing many highly similar genomes (differing by as little as 25 SNPs), allowing us to pinpoint specific strains to within 25 SNPs. This ability to pinpoint (mixes of) specific strains can aid in accurately identifying reinfection cases vs relapses, giving more useful results compared to the manual approach of the Bryant case-study with which our findings are consistent.

Using a systematic approach such as QuantTB aids in identifying cryptic transmission events, such as for samples with dissimilar major strains but matching minor strains. This may have occurred in two of the samples we surveyed in the data of Bryant et al. (samples 2 and 10). The ability to pinpoint strain mixtures can also aid in tracking progression of microevolution between sample isolates, including the evolution of resistance.

Using simulated data, we showed that QuantTB can accurately classify *M. tuberculosis* strains across a variety of database sizes. QuantTB is highly scalable, and can efficiently classify samples with databases as large as 4,000 strains in minutes, a necessary functionality as more and more TB assemblies are resolved. Other published tools made for classifying single strains in samples,

StrainSeeker and Sigma, were not capable of working with large databases, limiting their applicability as a diagnostic tool for *M. tuberculosis*. On tests using a smaller database - an easier and low resolution experiment - QuantTB identified the strain composition of synthetic sets with comparable accuracy as StrainSeeker, while Sigma's results included numerous false positives. On tests where the mixed samples contained strains absent from the database, QuantTB outperformed the other tools by accurately outputting the correct multiplicity in 72% of cases, in comparison to 25% for StrainSeeker and 0% for Sigma. Both Sigma and StrainSeeker consistently outputted aberrantly high number of strains. In addition, QuantTB predicted the closest related genome in the database for these strains in 94% of the samples.

The detection of high quality SNPs in a sample is an essential part of QuantTB's algorithm. In order to ensure erroneous SNPs are not considered, QuantTB disregards SNPs present at less than 5% abundance relative to that of the previously identified strain. Therefore, QuantTB can only detect mixed infections in which the minor strain represents at least 5% of the allelic variation. However, QuantTB is still able to pinpoint low-abundance strains with greater sensitivity than previous approaches based on the counting of heterozygous positions, due to its ability to identify strain down to coverages as low as 1x.

An advantage of approaches based purely on heterozygous locations is that they do not depend on a reference database. QuantTB's ability to accurately detect mixed infections is closely integrated with the distribution of genomes used to construct the database. Though we have tested QuantTB's performance on samples containing strains absent from the database, we have not extensively tested how the absence of a large proportion of a strain's lineage would affect its classification. QuantTB's ability to detect a strain not in the database depends on how distant it is from its nearest relative in the database. If the strain is sufficiently distant, it is likely that the strain would not be detected, underestimating sample diversity. The effects of QuantTB's database reliance is mitigated by ensuring the database covers as much diversity as possible. We found the currently available data is skewed to favor genomes of lineage 4 and lineage 2, with lineage 7 and 5 representing only 0.2% of the downloaded assemblies. Therefore, further sequencing of these underrepresented lineages would aid QuantTB in proper classification of novel strains.

QuantTB determines antibiotic resistance phenotypes by querying the sample against a manually curated list of SNPs that were shown to cause antibiotic resistances in previous studies. Bryant et al. did not find clinical evidence for antibiotic resistance amongst the samples. Using the curated list provided by Manson et al [24] we found antibiotic resistance in five samples, one being a case of heteroresistance in the second isolate of its sample pair. We did not observe any relationship between antibiotic resistance and mixed infections in the clinical isolates. The observed resistance mutations are well-known causal mutations for their respective resistances and WGS has been shown to outperform phenotypic susceptibility tests for predicting resistance [38]. Because Bryant et al. did not mention which method of genotypic testing they employed, it is not possible to understand why they were unable to detect genotypic resistances in the isolates. Particularly the *katG* mutation predicted from genotypic data in samples 49a and 49b is widely known and confirmed to confer resistance to isoniazid. The ability to accurately determine antibiotic resistance from sequencing data is still an active research topic for TB [39,40]. As antibiotic resistance is one of the biggest threats to world-wide TB eradication, the proper detection of possible resistance in samples is crucial.

2.7 Conclusion

We introduce QuantTB, a new classification method that leverages the high-resolution capability of whole genome sequencing for the detection of mixed *M. tuberculosis* infections. In contrast to existing tools such as Sigma and StrainSeeker, QuantTB is scalable and able to leverage a high-resolution reference database representing the scope of diversity within TB. Even when using a smaller database that allows comparisons between these tools, QuantTB shows substantially better performance on both synthetic and clinical datasets. This tool can be used to rapidly and accurately identify specific *M. tuberculosis* strains in clinical samples, track transmission of TB strains across longitudinal samples and outbreaks, and differentiate between relapse and reinfection cases. The ability to disentangle mixed infections in an accurate and scalable manner will help control TB and help limit the spread of antibiotic resistance.

2.8 References

1. World Health Organization. Tuberculosis Fact Sheet. 2018;

2. Colijn C, Cohen T, Murray M. Latent coinfection and the maintenance of strain diversity. Bull. Math. Biol. 2009; 71:247–63

3. Warren RM, Victor TC, Streicher EM, et al. Patients with Active Tuberculosis often Have Different Strains in the Same Sputum Specimen. Am. J. Respir. Crit. Care Med. 2004; 169:610–614

4. Cohen T, van Helden PD, Wilson D, et al. Mixed-strain Mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. Clin. Microbiol. Rev. 2012; 25:708–719

5. McIvor A, Koornhof H, Kana BD. Relapse, re-infection and mixed infections in tuberculosis disease. Pathog. Dis. 2017; 75:

6. Hatherell H-A, Colijn C, Stagg HR, et al. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. BMC Med. 2016; 14:21

7. Tarashi S, Fateh A, Mirsaeidi M, et al. Mixed infections in tuberculosis: The missing part in a puzzle. Tuberculosis 2017; 107:168–174

8. van Rie A, Victor TC, Richardson M, et al. Reinfection and Mixed Infection Cause Changing *Mycobacterium tuberculosis* Drug-Resistance Patterns. Am. J. Respir. Crit. Care Med. 2005; 172:636–642

9. Zetola NM, Shin SS, Tumedi KA, et al. Mixed Mycobacterium tuberculosis complex infections and false-negative results for rifampin resistance by GeneXpert MTB/RIF are associated with poor clinical outcomes. J. Clin. Microbiol. 2014; 52:2422–9

10. Richardson M, Carroll NM, Engelke E, et al. Multiple Mycobacterium tuberculosis strains in early cultures from patients in a high-incidence community setting. J. Clin. Microbiol. 2002; 40:2750–4

11. Dai S, Long Y. Genotyping analysis using an rflp assay. Methods Mol. Biol. 2015; 1245:91–99

12. Schürch AC, van Soolingen D. DNA fingerprinting of Mycobacterium tuberculosis: From phage typing to whole-genome sequencing. Infect. Genet. Evol. 2012; 12:602–609

13. Bryant JM, Harris SR, Parkhill J, et al. Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. Lancet Respir. Med. 2013; 1:786–792

14. Guerra-Assunção JA, Houben RMGJ, Crampin AC, et al. Recurrence due to Relapse or Reinfection With *Mycobacterium tuberculosis* : A Whole-Genome Sequencing Approach in a Large, Population-Based Cohort With a High HIV Infection Prevalence and Active Follow-up. J. Infect. Dis. 2015; 211:1154–1163

15. Sobkowiak B, Glynn JR, Houben RMGJ, et al. Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data. BMC Genomics 2018; 19:613

16. Gan M, Liu Q, Yang C, et al. Deep whole-genome sequencing to detect mixed infection of mycobacterium tuberculosis. PLoS One 2016; 11:1–14

17. Ahn TH, Chai J, Pan C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics 2015; 31:170–177

18. Roosaare M, Vaher M, Kaplinski L, et al. StrainSeeker : fast identification of bacterial strains from unassembled sequencing reads using user-provided guide trees . bioRxiv 2016; 1–20

19. Sankar A, Malone B, Bayliss S, et al. Bayesian identification of bacterial strains from sequencing data. bioRxiv 2015; 1–16

20. Hong C, Manimaran S, Shen Y, et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2014; 2:33

21. Shin SS, Modongo C, Baik Y, et al. Mixed Mycobacterium tuberculosis-Strain Infections Are Associated With Poor Treatment Outcomes Among Patients With Newly Diagnosed Tuberculosis, Independent of Pretreatment Heteroresistance. J. Infect. Dis. 2018;

22. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016; 44:D733–D745

23. Benson DA, Karsch-Mizrachi I, Lipman DJ, et al. GenBank. Nucleic Acids Res. 2011; 39:D32–D37

24. Manson AL, Cohen KA, Abeel T, et al. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. Nat. Genet. 2017; 49:395–402

25. Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes. 2004; 5:

26. Walker BJ, Abeel T, Shea T, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One 2014; 9:e112963

27. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;

28. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078–2079

29. Clark TG, Mallard K, Coll F, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. PLoS One 2013; 8:e83012

30. Crispell J, Zadoks RN, Harris SR, et al. Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand. BMC Genomics 2017; 18:180

31. Sampson SL. Mycobacterial PE/PPE Proteins at the Host-Pathogen Interface. Clin. Dev. Immunol. 2011; 2011:1–11

32. Albanese D, Donati C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat. Commun. 2017; 8:2260

33. Huang W, Li L, Myers JR, et al. ART: A next-generation sequencing read simulator. Bioinformatics 2012; 28:593–594

34. Leinonen R, Sugawara H, Shumway M, et al. The sequence read archive. Nucleic Acids Res. 2011; 39:D19-21

35. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 2009; 26:1641–50

36. Coscolla M, Gagneux S. Consequences of genomic diversity in mycobacterium tuberculosis. Semin. Immunol. 2014; 26:431–444

37. Pérez-Lago L, Comas I, Navarro Y, et al. Whole Genome Sequencing Analysis of Intrapatient Microevolution in Mycobacterium tuberculosis: Potential Impact on the Inference of Tuberculosis Transmission. J. Infect. Dis. 2014; 209:98–108

38. Feliciano CS, Namburete EI, Rodrigues Plaça J, et al. Accuracy of whole genome sequencing versus phenotypic (MGIT) and commercial molecular tests for detection of drug-resistant Mycobacterium tuberculosis isolated from patients in Brazil and Mozambique. Tuberculosis 2018; 110:59–67

39. Bradley P, Gordon NC, Walker TM, et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nat. Commun. 2015; 6:10063

40. Walker TM, Kohl TA, Omar S V, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect. Dis. 2015; 15:1193–1202

3 Whole genome sequencing reveals extensive prevalence of mixed tuberculosis infections across the globe and its role in drug heteroresistance.

3.1 Abstract

Mixed infections of *Mycobacterium tuberculosis* have been associated with poor treatment outcomes for tuberculosis; however, their prevalence and impact have been rarely studied, which is partly attributed to limitations in their detection. The primary objective of this study was to quantify the degree of mixed TB infections using publicly available *M. tuberculosis* DNA sequencing data in order to characterize their impact on studies of TB epidemiology and drug resistance.

In analyzing a collection of 48,392 *M. tuberculosis* DNA sequencing read sets from greater than 100 studies conducted in 106 countries, we detected mixtures of *M. tuberculosis* isolates in 6% of samples. These mixtures were distributed across 80% of included studies and 95% of countries indicating the widespread nature of mixed TB infections. Mixed infections were more likely to occur in strains within the same lineage, which would likely be missed using common DNA fingerprinting techniques. Importantly, mixed infections accounted for almost a third of strains where we detected heteroresistant loci, or cases where both susceptible- and resistance-conferring alleles were detected. In most cases, heteroresistance could be explained by the presence of multiple infecting strains within the sample as opposed to within host evolution of a resistant genotype.

3.2 Introduction

Tuberculosis (TB) is one of the deadliest infections in the world, with greater than 10 million active cases diagnosed in 2018 [1]. Traditionally, TB disease was thought to represent a single infecting strain of *Mycobacterium tuberculosis* and the possibility of infection due to more than one isolate was rarely considered [2]. However, advances in molecular techniques have identified numerous cases in which individuals harbor multiple *M. tuberculosis* strains, a phenomenon known as mixed *M. tuberculosis* (TB) infections [3]. Mixed TB infections have been increasingly reported, with prevalence estimates varying between 0 and 20% in small scale studies [4–7], with host-immune status playing a role [6]. In some cases, a person can be infected with strains exhibiting differing resistance phenotypes [8–10]. Numerous studies have shown mixed TB infections often lead to poor treatment outcomes, diminished performance of diagnostic tools, and reduced treatment efficacy [11–15], which highlights the need for additional investigation of mixed TB infections.

Difficulties in identifying mixed infections have led to limited understanding of their prevalence and impact. Initially, DNA fingerprinting techniques like IS6110 restriction fragment length polymorphism (IS6110 RFLP) [16], spacer oligonucleotide typing (spoligotyping) [17], and

analysis of mycobacterial interspersed repetitive-unit-variable number of tandem repeats (MIRU-VNTR) [17] were among the most popular methods to differentiate *M. tuberculosis* lineages. Though these methods were also used to detect mixed infections, they have low discriminatory power to differentiate between highly similar strains within a lineage [3]. The advent of whole-genome sequencing (WGS), which provides a high resolution view of the diversity among *M. tuberculosis* strains present in a sample, allowed for substantial improvement in the discovery of mixed infections relative to DNA fingerprinting techniques [3,18]. Although WGS has become fundamental for the study of TB epidemiology, antibiotic resistance, transmission, and diagnostics [19,20][21,22][23], the assumption that TB disease is typically the result of a single infecting strain has meant that many large-scale WGS analyses have not incorporated the importance or influence of mixes on their research process or their research results. Recently, computational methods to identify mixed TB infections within WGS data [24–26] have allowed researchers to identify mixed infections with higher resolution from WGS data sets. However, a large-scale, systematic study specifically focusing on understanding the global prevalence of mixed infections in TB has yet to be performed.

The primary objective of this study was to explore the prevalence and nature of mixed TB infections within more than 50,000 publicly available TB WGS sequences and their impact on TB research representing every major global geographical region. We investigate the prevalence of mixed infections across geographical regions, characterize properties of a mixed infection with respect to lineage and strain multiplicity, and evaluate the relationship of mixed infections to drug resistance, heteroresistance and phylogenetic analyses. We observe that mixed infections were present within every geographical region, with particularly high identification rates in countries of the Caucasus region. These mixed infections comprised strains from every TB lineage – with overrepresentation of lineage 2 within mixed samples. We identified mixed infections containing more than two strains which we denoted as ‘supermixes’. Lastly, we report a high frequency of drug resistance within mixed samples, particularly heteroresistance.

3.3 Materials and methods

3.3.1 Dataset construction and variant identification

53,025 WGS short-read datasets were downloaded from NCBI’s sequencing read archive (SRA) [27] on December 12, 2018 using the taxonomic ID for *M. tuberculosis* (txid1494075). After retaining only samples sequenced on the Illumina platform, with a read length >= 50 bp and coverage >= 20x, our final working dataset contained 48,392 samples. The full list of SRA file identifiers and their BioProject accessions from our final dataset can be found in Supplementary Table 1.

SRA files were extracted using fastqdump (Version 2.9.0) [27] from the SRA toolkit, using the “split-3”, “skip-technical”, and “clip” flags to split left and right reads into separate files, remove technical reads, and clip poor-quality ends of reads, respectively. The extracted read sets were aligned against the *M. tuberculosis* H37Rv reference genome (Genbank identifier CP003248.2) with BWA-MEM version 0.7.17-r1188 [28] using default settings, then index-sorted with samtools version 1.6 , using htslib 1.6 [29]. Due to the wide range of coverages observed within the read

sets, each aligned bam file was down-sampled to a maximum coverage of 150x, as in our previous large-scale analysis of *M. tuberculosis* genomes [20].

As described previously [20], we assigned lineages to each read set based on lineage specific markers and used our alignments against the H37Rv reference to call variants using Pilon (version 1.22, default settings, fixes set to none) [30]. Variants located in genes annotated as PE/PPE within the H37Rv reference were excluded (Supplementary Table 2), as these genes are known to be highly repetitive and prone to mapping errors, making it difficult to call variants using short-read data [31–33].

The geographic region of isolation for 70% of all SRA samples could be determined based on their BioSample metadata identifier (37% of samples) or from their corresponding publication (an additional 33% of samples), which we examined for studies with at least 100 samples. We also tabulated the goals of each study using a set of 12 keywords based on their abstracts for studies with at least 100 samples (Supplementary Table 3), and determined whether each study checked for mixed infections. Where possible, we tabulated the method by which samples were processed, which included 40 samples that were isolated from a plate sweep from a previous study of ours [19], and 1,612 samples that represented single colonies isolated from 5 studies [19,34–37].

3.3.2 Identifying mixed infections

To assess whether a sample contained greater than one infecting *M. tuberculosis* strain, we used QuantTB [24], a tool we recently published which accurately and precisely identifies mixed infections of *M. tuberculosis* strains that differ by as few as 25 SNPs. We previously demonstrated that QuantTB outperformed other existing tools [38–40] for identifying mixed infections in *M tuberculosis* [24]. In brief, QuantTB determines mixed infections by iteratively comparing a database of curated reference genomes against every sample to identify strains present based on common allelic distribution. QuantTB was run for each sample on VCF files containing variant calls generated by Pilon.

As the ability of QuantTB to detect strains in a sample is highly dependent on the representation of close matches to the strains within the reference database, we established a comprehensive set of reference databases to ensure that we could efficiently cover the diversity within our large sample set. These databases included: (i) the default database provided by QuantTB, which contains 2,162 *M. tuberculosis* genomes from RefSeq (Supplementary Table 4) separated by at least 100 SNPs, and (ii) a more inclusive database of 3337 *M. tuberculosis* genomes, separated by at least 50 SNPs. We constructed this second database from our sample read sets by selecting samples predicted to represent single-isolate infections only, using a previously published method to identify and exclude mixed infections based purely on the presence of heterozygous sites [41]. We selected samples from our dataset with < 0.0001% heterozygous SNPs (< 5 heterozygous SNPs genome–wide).

After running QuantTB using each of these two independent databases, we merged the results to create the final prediction set. We further assessed our samples for the presence of false

negatives due to insufficient coverage of the reference database. If QuantTB predicted a single strain, but estimated that it accounted for <80% of the heterozygosity in the sample, it is likely that there was actually another strain present, which QuantTB was unable to identify because there was no reference genome similar enough in the reference database. For such cases, the sample was classified as a false negative and labeled as a mix.

3.3.3 Prediction of genotypic antibiotic resistance

Genotypic drug resistance was identified using QuantTB. Resistance conferring alleles were defined from a curated set of SNPs conferring resistance to seven anti-tubercular drugs [20]. A sample was considered homoresistant, or fully resistant to a particular drug, if at least one resistance-conferring allele was detected at a frequency of > 90%. A sample was considered heteroresistant when it harbored alleles for both resistant and susceptible genotypes, and at least one resistance-conferring allele was detected at a frequency between 10-90%. A sample was considered drug susceptible if it contained no drug resistance-conferring alleles at frequency of at least 10%.

In order to identify cases where heteroresistance may have arisen through mixed infections of strains harboring different drug susceptibility profiles, we associated each strain in a heteroresistant mixed infection to the observed drug susceptibility profile. For each heteroresistant allele, the abundance levels of the susceptible alleles and resistant alleles were matched to a component strain based on the overall strain frequencies output by QuantTB if the frequencies were within 0.1 from each other. Only cases with a proper match are reported.

3.3.4 Phylogenetic tree construction

To construct a phylogenetic tree, we used an alignment of genomic positions containing SNPs identified by Pilon. Insertions, deletions, bases with low quality (Phred score less than 11) and positions located within PE/PPE regions were not considered. FastTree [42] was used to generate a tree from the remaining SNPs with default settings.

3.4 Results

3.4.1 Mixed TB infections are present in 6% of sequenced samples globally

To quantify the global prevalence of mixed TB infections, we used QuantTB, a recently published, publicly available tool to identify mixed infections in a dataset of 48,392 *M. tuberculosis* short-read datasets from NCBI. Using default thresholds, QuantTB identified mixed infections—samples containing *M. tuberculosis* genomes from more than one infecting isolate--in 6% of samples across 149 BioProjects. Using QuantTB with WGS data, we were able to detect five times as many mixed infections compared to the standard sequencing-based molecular fingerprinting approach, which we implemented on our sample set by identifying the presence of multiple lineage markers in samples using the *macaw* algorithm[20].

Overall, mixed infections were identified in 80% of the 149 BioProjects surveyed, highlighting the prevalence of mixed infections globally and in every domain of TB research. Mixed infections seemed to arise in nearly all settings that have been analyzed with WGS, despite the fact our dataset was pooled from a large collection of datasets, representing different facilities and protocols for mycobacterial sample processing. The observed rates of mixed infection per study varied from 0% to 54% (Supplementary Figure 1). This large range of rates across the studies could be due to many factors, including study protocol, study objectives, and study population. Indeed, we observed studies with the highest levels (30+%) of mixed infections represented studies conducted in countries with high TB incidence, such as Iran (PRJNA237443) and Ethiopia (PRJEB9201), suggesting a some link between TB incidence and the rate of mixed infections. However, plotting WHO reported TB incidence rates vs. detected mixed infection rates does not imply a nice linear relationship.

To evaluate if application of differing microbiologic techniques upstream of sequencing could confound our ability to detect mixed infections among WGS data, we separately analyzed the mixed infection rate among samples that were single colony isolated prior to sequencing (1,612 samples) vs. confirmed plate sweep (40 samples) (see Methods). Irrespective of pre-sequencing microbiologic techniques, we observed the same rate of mixed infections (3%) among these differently processed samples. This observation suggests that mixed infections can still be identified regardless of the method used to isolate samples, and that despite best efforts to obtain a pure colony, single colony isolation techniques do not always yield a genetically homogeneous population during sequencing.

Due to difficulties with disambiguation, the widespread presence of mixed infection in TB has been largely ignored, and sometimes mixed infections have been excluded from analyses. Publications for only 20 (27%) of the 74 Bioprojects with > 100 samples indicated they had a process to screen for mixed infections in their quality control process, such as by counting the number of bi-allelic sites or quantifying heterozygosity.

In order to understand more about the kinds of studies associated with different levels of mixed infections, we ascertained the objective of each study based on abstract review, and grouped them into 12 categories of study objectives. Irrespective of study objective, mixed infections were observed across all types of TB research, including investigation of resistance, transmission, or diagnostics (see Methods; Supplementary Table 3). We observed that studies focused on drug resistance/prediction contained the highest amount of mixed TB infections, while the highest proportion of mixed infections (12.6%) was observed in studies focusing on recurrence of TB.

3.4.2 Mixed TB infections observed on all continents

We performed a more in-depth analysis of geography and mixed TB infection rates for a subset of 33,641 samples (70% of our dataset) for which the country of isolation was known. 106 countries and territories from around the globe were represented. Among those countries contributing at least 20 samples to our dataset, nearly all (50/51) were found to have examples of mixed TB infections. The one exception was Ivory Coast (see Figure 3.1).

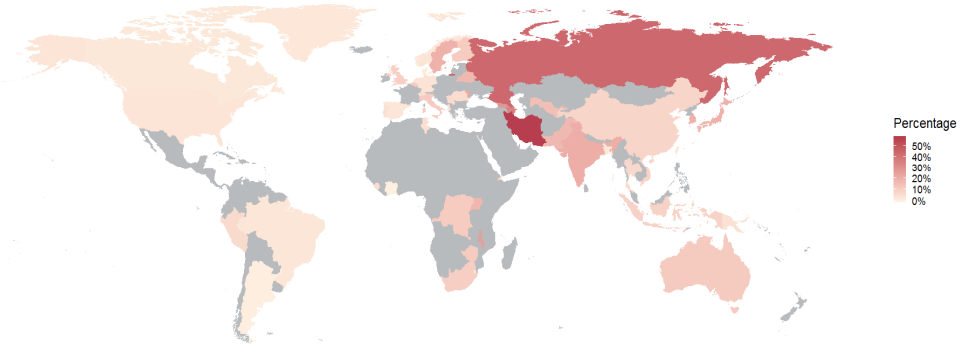


Figure 3.1 Percentage of mixed TB infections identified in samples from the 51 countries with at least 20 samples in our data set. A lighter color denotes a lower percentage of mixed infections; darker red denotes a higher percentage of mixed infections. Grey indicates insufficient data.

In agreement with our hypothesis, countries having low incidence of TB had some of the lowest rates of mixed TB infections, which may be due to the low likelihood of encountering multiple strains. Very low incidence countries have, as expected, the lowest rates of mixed TB infections, i.e. the Americas (<3%) and Western Europe (<2%). Because these regions have very low incidence of TB, it is not likely that a patient would have been infected with multiple strains. However, based on our data set, countries with high TB incidence, such as South Africa, have an average mixed infection rate of 3%. Surprisingly, it is a set of moderate-high burden countries around the Caucasus that were associated with higher mixed TB rates, including Iran (57%), Russia (43%), Azerbaijan (35%), and Georgia (27%) (Figure 3.1, Supplementary Table 4). Among the higher burden TB countries in our dataset, we observed moderate rates of mixed infections (<25%) in African and South Eastern nations.

3.4.3 The phylogenetic distribution of mixed infections differs from that of single isolates

Due to human-to-human transmission of TB, we would expect the distribution *of M. tuberculosis* lineages within mixed infections to mirror that of single infections globally. While the distribution of lineages across single isolates in our >48,000 sample dataset is similar to those reported in previous work [43], the phylogenetic distribution of mixed infections differed significantly from that of single isolates ((Figure 3.2, Supplementary Figure 2). For instance, lineage 4 strains were significantly less commonly observed within mixed infections (32%) than single isolate infections (52%) (p-value < 3e-119) (Figure 3.2). In contrast, lineage 2 strains were present within mixed

infections significantly more often than expected (p-value < 4.6e-73). Despite being relatively small contributors to global TB burden, lineages 3 and 6 were also disproportionately overrepresented within mixed TB cases (p-values < 1.3 e-10 and <8.87 e-27 respectively).

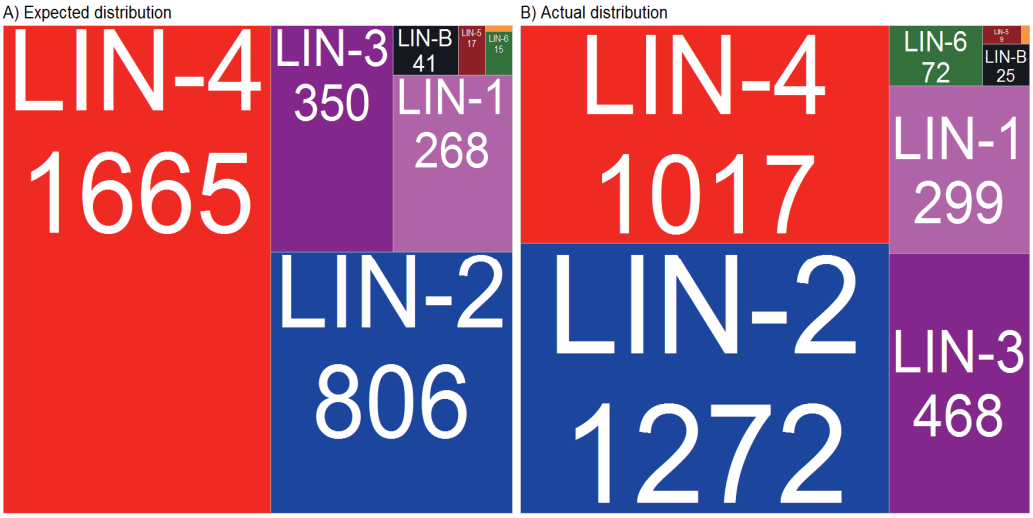


Figure 3.2 The global distribution of M. tuberculosis lineages within mixed samples (right) differs from that expected based on single isolates (left) within our dataset of 48,392 strains. The expected values are calculated by rescaling the distribution values of the distribution of lineages of isolates in the complete data set to match the number of observed mixed samples.

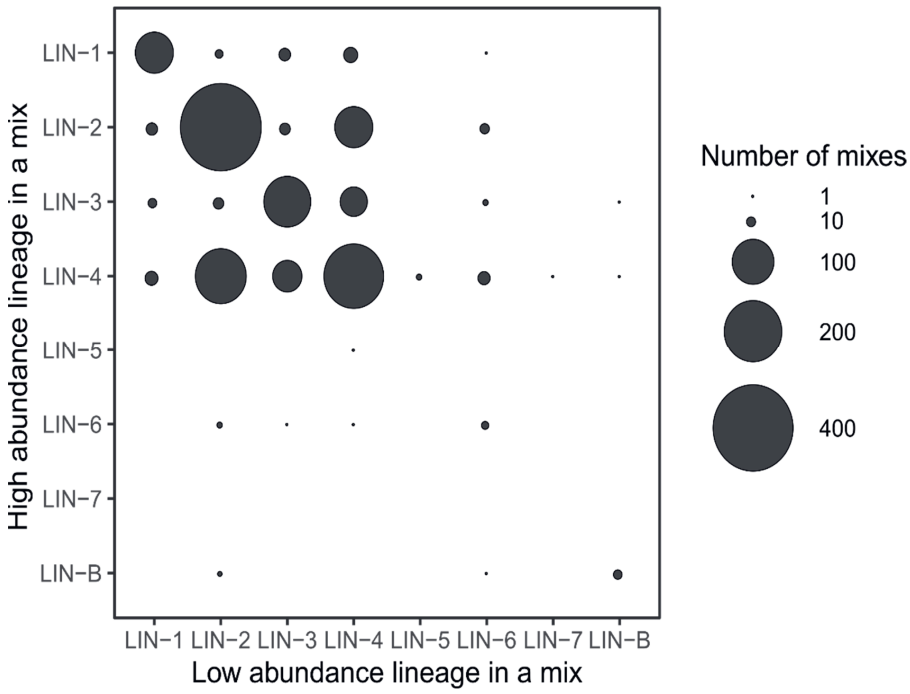


Figure 3.3 Most mixed infections involved strains from the same lineage, with mixed infections between lineage 2 strains being the most abundant. Correlation plot of the different lineages present within mixed infection pairs, with the higher abundant strain in the pair on the x axis and the lesser abundant strain in the pair on the y axis.

Furthermore, we found that most mixed infections (62%) involved strains from the same lineage, which may not be surprising given the predominance of some lineages with some geographies. Mixed infections involving two lineage 2 strains were the most frequent (Figure 3.3), and the majority of mixed infections between lineages occurred between lineage 2 and lineage 4. Strains from Lineage 6, one of the *M. africanum* lineages, were more frequently observed in mixed infections than expected from their occurrence in single infections (Figure 3.2) but they were almost always the lesser abundant strain of a mixed pair, which could possibly represent differences in virulence, as lineage 6 is known to be less virulent [44–46].

3.4.4 3.5% of mixed TB infections are “supermixes”, or mixtures of 3 or more strains

Likely due to technical difficulties in identifying mixes, only a limited number of cases with greater than two TB strains have been reported within a single sample [4]. We defined “supermixes” as samples that contained mixtures of three or more unique TB infection strains. Using QuantTB, we found that such supermixes accounted for 3.5% of all mixed infections, or 103 isolates (0.2% of all isolates). The majority of these (95%) contained three strains; only four instances contained

four strains. As was seen overall, lineage 2 strains were the most likely to be present within a supermix. Also mirroring all mixtures, supermix samples were identified across many different countries (21) and Bioprojects (36), and were not enriched within any geography or study.

3.4.5 Mixed TB infections lead to improperly classified transmission events and missed connections between samples

We constructed a phylogenetic tree (Supplementary Figure 3) using concatenated SNPs (see Methods) and used it to compare the placement of samples predicted by QuantTB to represent either a mixed infection of two strains or single isolates of the same two strains. We aimed to identify whether mixed infections clustered in a phylogenetic tree, or whether they were placed in the neighborhood of samples predicted to contain one of their constituent strains. This view serves as a useful tool understanding how mixed infections may have caused phylogenetic misinterpretations in prior literature, when they were interpreted as single isolates (Figure 3.4).

Twenty specific combinations of references occurred five or more times in our dataset. By examining subsets of the phylogenetic tree corresponding to individual isolates or mixes of just these combinations of references (Figure 3.4, Supplementary Figures 4-5), we observed that reference combinations appeared primarily next to individual-isolate samples identified as most similar to one of the constituent references (usually the higher-abundance strain) (Figure 3.4, Supplementary Figure 4), or primarily in a separate location specific to that mixed genotype (Supplementary Figure 5). In all but one case, the location in the tree of a sample with a specific mix of references was closer to samples which contained the most abundant reference within that mix.

These phylogenetic data reveal that samples containing undetected mixed infections can be placed in surprising and misleading locations on the phylogeny, which can be uninformative as to the nature of the component strains. Interestingly, while some of these mixes of specific references derived from multiple specimens from the same patient, the majority of instances derived from different patients or different investigations, indicating likely co-transmission of the pair of strains or local differences in strain prevalence of two infecting strains.

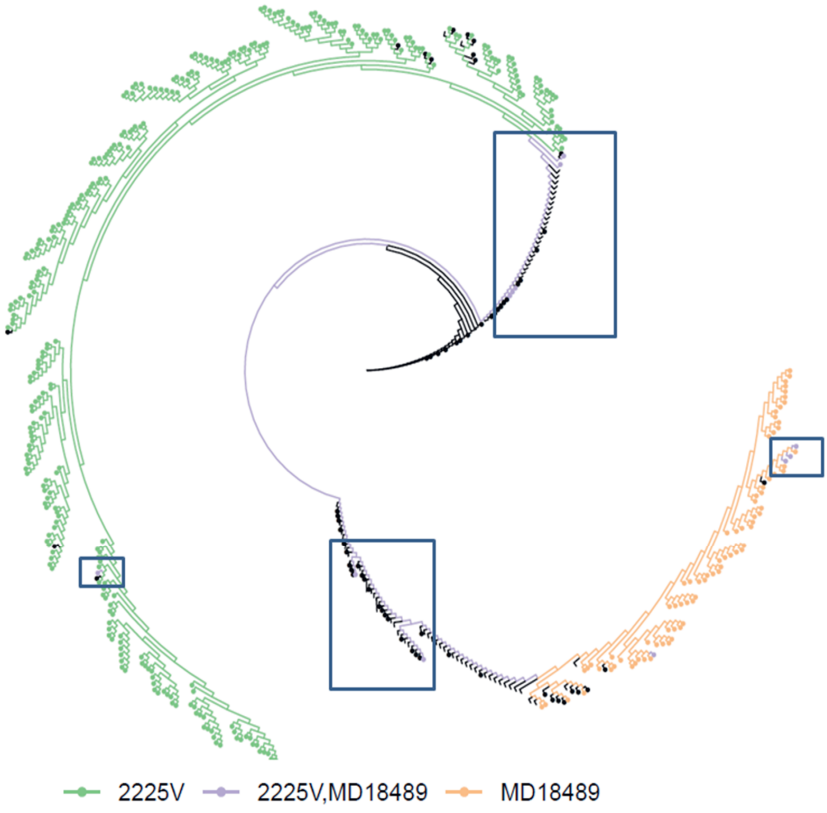


Figure 3.4 Example of differing placement of mixed infection pairs. This tree shows all the local regions of a phylogenetic tree highlighting with single isolate samples of strain 2225V (green) or MD1849 (orange), together with samples containing a mixture of these two strains (purple). The tree tips and nodes that did not pertain to these 3 strains were collapsed to simplify visualization. Boxes indicate clusters of mixed samples, i.e. purple leaf nodes. The small left-most box indicates 1mixed sample placed near strain 2225V; the right-most box indicates 3 mixed samples placed near strain MD18489; and the two larger central boxes indicate 7 mixed samples that are not placed near either of the constituent strains.

3.4.6 Mixed infections impact antibiotic resistance

Based on a curated list of SNPs for resistance genes [20], we found that mixed infections had a higher prevalence of resistance than single-isolate samples (see Methods). There was a smaller but significantly higher prevalence of resistance within mixed infection samples (42% of samples) than in single-isolate samples (38%) (p-value < 4e-7), including for multi-drug resistance (MDR) (19% vs. 17%) and extensively drug resistance (XDR) (4% vs. 2.5%). Overall, mixed infections accounted for 7% (565) and 10% (120) of all MDR and XDR samples, whereas mixed infections represented only 6% of all samples.

Furthermore, recent studies have shown that mixed infections can result in heteroresistance, a phenomenon in which samples contain both resistant and susceptible genotypes, which can complicate TB resistance classification and treatment recommendations. Although the majority of resistance for all antitubercular drugs analyzed was due to a single resistance allele, heteroresistance was present for every drug, at rates ranging from 3% of isoniazid-resistant samples to 33% of ethionamide-resistant samples (Figure 3.5, panel A). Heterozygous-resistance, where multiple distinct mutations were detected that can cause resistance to the same antibiotic, was rarer (<1% of samples) and detected for only four drugs (ethambutol, isoniazid, ofloxacin, and rifampicin). Strikingly, mixed samples accounted for almost a third of strains where we detected heteroresistant loci (average of 31%; with per-drug rates ranging from 17.5% of heteroresistance to ofloxacin, to 56% for isoniazid) (Figure 3.5, panel B). The remaining two-thirds of heteroresistance, found within single-isolate samples, was most likely caused by within-patient microevolution.

In order to quantify strains contributing to heteroresistance amongst mixed infections, we identified which strain in a mixed sample was responsible for the observed heteroresistance. We matched the allele frequency at each heteroresistance conferring locus to the corresponding abundance levels of strains in the mixed sample and determined whether or not the heteroresistance was due to either 1) a mixture of two strains with distinct drug resistance profiles or 2) a mixture of a resistant strain and a susceptible strain.

We were able to match the resistant allele frequency to the abundance of a strain in the mixture for 416 of the 628 heteroresistant mixed samples (66%%) (see Methods). Of these, we determined that the majority of heteroresistant mixed infections (74%) were attributed to two infecting strains with distinct drug resistance profiles. For example, sample ERR2679289 contained a mix of two infecting strains (83% majority strain vs. 17% minority strain) and was predicted to be MDR. However, upon further analysis, the MDR genotype was attributed to an isoniazid mono-resistant (80% resistant allele) infection from the majority strain, and a rifampicin mono-resistant (11% resistant allele) minority strain. As each of the two infecting strains were in fact mono-resistant to the drugs that define of MDR-TB, this case highlights how heteroresistance can confound the diagnosis of MDR, as this sample may have been incorrectly identified as MDR on culture-based phenotypic drug susceptibility testing or on molecular resistance testing. For a large minority of cases (26%), the heteroresistance analysis revealed the sample contained a mixture of a drug resistant strain together with a fully susceptible strain. For example, sample ERR484733, an MDR sample which was heteroresistant to rifampicin, isoniazid, and streptomycin (with 17%, 20%, and 11% resistant allele frequencies, respectively) was determined to represent a mixture of a fully susceptible strain at estimated 75% abundance, and strain resistant to all three drugs, at an estimated 25% abundance. This case represents an example of where an undetected resistance due to a low minority strain could lead to treatment failure or further development of resistance. For both of these examples, knowledge that mixed infection was contributing to the observed resistance genotypes could have been exploited for optimal design of a TB treatment regimen.

Finally, samples with “supermixes” harbored increased rates of drug resistance (62% of samples) and heteroresistance (52% of samples) compared to single isolate infections (37% and 7% respectively) and mixes of only two strains (42% and 24% respectively).

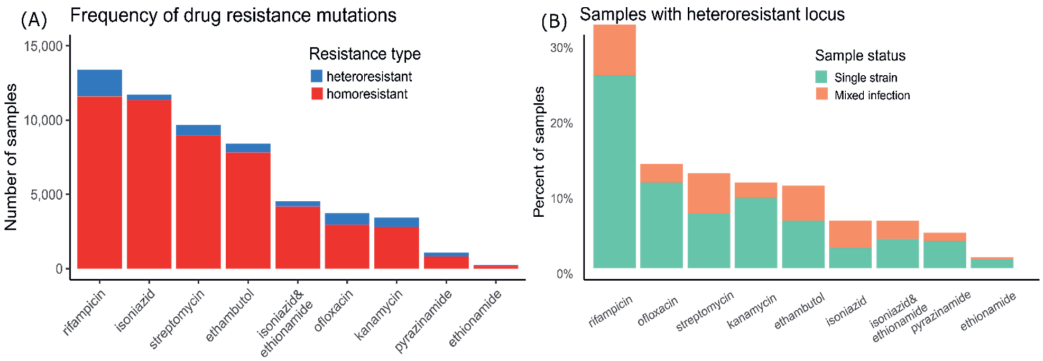


Figure 3.5 Panel A shows heteroresistance all drugs at varying rates. The bars represent the number of samples containing at least one resistant conferring mutation. Samples were categorized as i) heteroresistant (orange; susceptible and resistant alleles both present) and homoresistant (green; only a single resistant conferring allele is present). Panel B showns the the number of samples with at least one heteroresistant locus for each drug. Mixed infections were highly enriched among heteroresistant samples (overall average of 31% of heteroresistant samples, as compared to 6% of all samples). Samples with mixed infections were found in almost equal proportions across all the drugs surveyed.

3.5 Discussion

To the best of our knowledge, this study portrays the broadest examination of mixed infections to date, leveraging the largest and most comprehensive set of *M. tuberculosis* WGS data from studies performed across the globe. This work provides new insight into the rate of mixed infections and its impact on heteroresistance. Our classification method identified 5 times more mixed infections compared to a sequencing based molecular fingerprinting based detection method, implying that rates based on spoligotyping or other molecular markers underestimate the amount of mixed infections.

Across our large collection of datasets, each sourced for a different study goal and application, approximately 6% of samples were identified to contain mixed TB infections. Mixed infections were identified in nearly all countries' samples, irrespective of the type of investigation or analysis. Highlighting the widespread prevalence of mixed TB infections fills a previous knowledge gap, as the majority of prior genomic studies purposefully excluded or did not attempt to identify the presence of mixed infection. Previous studies identified higher rates of mixed infections in countries with high TB incidence, presumably due to the increased occurrence of interpatient contact [47]. The differences in mixed infection rates between previously published results and our own could potentially be explained by low sampling across the African continent in our dataset, and indicating a need for additional TB genomic studies in these underrepresented geographies.

Relative to other *M. tuberculosis* lineages, lineage 2 strains contributed disproportionately to mixed infections. The observation that lineage 2 strains were more likely to develop mixed infections than lineage 4 strains is consistent with prior reports of the global dominance of lineage 2, [48,49] This over-representation could be explained by the fact that lineage 2 strains are thought to be more virulent and infectious [50,51]although further studies are needed to

understand why lineage 2 strains are more likely to coinfect. In addition, this study showcases how mixed TB infections can confound accurate phylogenetic analysis. Phylogenetic placement of samples containing mixed infections varied based on the abundance levels of the constituent strains. As phylogenetic methods are the basis for many key analyses, including dating methods, transmission profiling [23] and establishing the order of acquisition of drug resistance mutations [20], such errors in tree topologies have far-reaching consequences.

We showed that mixed infections are also a strong contributor to heteroresistance, or samples observed to contain both resistant and susceptible genotypes. A burgeoning area of interest in TB [52–54], heteroresistance may lead to failures in resistance detection or resistance profiling, leading to treatment failure. For example, if a patient harbored a mixed infection consisting of both rifampicin mono-resistant and isoniazid mono-resistant isolates, either phenotypic testing or molecular diagnostics (GeneXpert or Hain line probe assays) would detect the presence of both rifampicin and isoniazid resistance, and the patient would be assumed to harbor MDR-TB. In this instance, proper mixed status detection, could lead to more accurate treatment as the patient could be treated with the drugs necessary to kill the two monoresistant strains. Alternately, if there was a mixed infection with both a fully drug susceptible isolate and one or more isolates with additional drug resistances, the presence of the susceptible isolate would likely go unnoticed. Thus, the lack of detection of mixed infections with varying resistance profiles may lead to suboptimal drug regimen selection for patients. If resistance to both rifampin and isoniazid was detected, current WHO TB treatment guidelines [55] would not include rifampicin or isoniazid. However, if it was known that a mixed infection was present and included monodrug-resistant or susceptible strains, then both of these first-line TB drugs would have been included in the treatment regimen. Thus, undetected mixed infections may result in withholding key first-line drugs from TB treatment regimens, resulting in delayed bacterial clearance and increased toxicity.

While heteroresistance has often been attributed to mixed infections, the contribution of mixed infections toward heteroresistance has been evaluated only in limited datasets [56]. In this large global study, the fraction of heteroresistance attributable to mixed infections ranged from 13% for rifampicin to 57% of isoniazid heteroresistance. For the majority of TB drugs, heteroresistance was largely due to a single infecting isolate with heterogeneity only at a specific drug resistant locus, rather than mixed infection. The clinical importance of heteroresistance, which may represent emerging resistance evolution, remains an area of active investigation.

Although previous studies have examined strain mixes, this is the first study to have reported multiple cases of “supermixes” with greater than two *M. tuberculosis* isolates, made possible by the high resolution with which QuantTB can detect and quantify samples harboring mixed infection. Supermixes comprised 0.36% of all samples in our study, showed a strong association with heteroresistance, and suggested that within-patient heterogeneity is likely more complex than previously assumed. The relationship between identification of supermixes and clinical outcomes is not known; however, this phenomenon warrants further exploration as it may contribute to TB treatment failure.

As this investigation brought together a large collection of publicly available TB genomic specimens from diverse sources, meta-data was not uniformly available across samples,

including sampling methods, geographic locations, treatment outcomes, and isolation techniques. For example, we could not identify the geographical location for 30% of the samples in our dataset, potentially resulting in unintentional bias towards studies which provided this information. Additionally, specimen processing methods prior to sequencing were not clearly reported for many studies or varied widely. While a subset of studies used methods that would allow for broader detection of heterogeneity, such as plate sweeps, indirect sequencing from mycobacterial growth indicator tubes (MGIT), or direct sequencing, the remainder of studies used single colony isolation to obtain a homogeneous population. This may have eliminated or reduced the diversity of mixed TB infections within sequenced samples, resulting in an underrepresentation of mixed infections in our dataset. Although we reported similar mixed infection rates from single colony isolation or plate-scrape samples, this comparison was based on a limited number of samples for which this information was reliably available. It would be valuable to perform a direct comparison of the rates of mixed infections across different isolation techniques for a larger set of samples with known isolation techniques. Information on clinical outcomes was also not consistently present in our metadata. As this information would have been useful to obtain a deeper understanding of the clinical impact of mixed infections, future practices for documenting information within public databases would benefit from incorporating clinical outcomes.

Although future initiatives to perform more comprehensive sampling of the global *M. tuberculosis* population will provide a more complete view of mixed infections in TB, our current dataset was limited in a number of ways. First, the studies surveyed all had different objectives which impacted sample selection for sequencing. For example, certain studies restricted their patient population to individuals with phenotypic drug resistant strains, or TB-HIV co-infection, or were limited to a particular geographic region. As this collection represents a medley of different sequencing strategies and patient populations that may not accurately reflect global TB epidemiology. Due to our conservative approach to identify mixed infections, our results are likely an underestimate of the true incidence of mixed infections.

Second, high burden TB countries with resource limitations were inadequately represented in our dataset, which could lead to a bias in our understanding of mixed infections. Of the eight countries accounting for 66% of new cases in 2018 (Pakistan, Nigeria, Indonesia, the Philippines, Bangladesh, South Africa, India, and China [1]), only South Africa and China (4519 and 1237 samples respectively) had significant representation in our dataset. The relative novelty and expense of whole genome sequencing could be responsible for the fact that there were fewer NGS studies in part of the world with heavier disease burden. In addition to bias in the types of populations studied, this also resulted in bias in the composition of *M. tuberculosis* lineages in our dataset, because of geographic differences in lineage distribution.

Finally, Lineages 6 and 7 were severely underrepresented in our dataset, with only 192 and 28 samples in our dataset, respectively. This distribution was expected based on the global distribution of lineages, where lineages 5 and 6 are age generally geographically restricted to Western Africa and lineage 7 to the Horn of Africa. However a better understanding of these understudied lineages could yield a better understanding of *M. tuberculosis* evolution and provide a more comprehensive picture of mixed TB infections.

3.6 Conclusion

We identified mixed infections in 6% of WGS samples examined, across a global dataset containing over 48,000 samples. Mixed infections were present in virtually every country surveyed, with highest abundance in high-TB burden countries surrounding the Caucasus, including Russia, Iran and Azerbaijan. Mixed infections were more likely to occur between two strains from the same lineage, which is likely attributable to the proximity of lineages in local environments. Lineage 2 and lineage 6 strains were overrepresented within mixed infections, which could point to their increased ability to infect a host. Mixed infections were found in more than 80% of the biological studies surveyed, but were more prevalent in studies focusing on drug resistance, likely due to the strong association between mixed infections and heteroresistance, which is also widespread. Our results highlighted the impact of mixed infections on the placement of strains within phylogenies, which can substantially disrupt our ability to conduct transmission analysis and trace the origins of drug resistance. Lastly, we identified multiple circumstances in which lack of detection of mixed TB infection may hamper efforts to both diagnose and properly treat of drug-resistant TB. This study highlights the importance of identifying mixed TB infections within epidemiological and clinical studies that use WGS to characterize clinical isolates.

3.7 References

1. Global Report 2019.

2. Stead WW. Pathogenesis of a first episode of chronic pulmonary tuberculosis in man: recrudescence of residuals of the primary infection or exogenous reinfection? Am. Rev. Respir. Dis. 1967; 95:729–745

3. Cohen T, van Helden PD, Wilson D, et al. Mixed-strain Mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. Clin. Microbiol. Rev. 2012; 25:708–719

4. Das S, Narayanan S, Hari L, et al. Simultaneous infection with multiple strains of Mycobacterium tuberculosis identified by restriction fragment length polymorphism analysis. ingentaconnect.com

5. Chaves F, Dronda F, Alonso-Sanz M, et al. Evidence of exogenous reinfection and mixed infection with more than one strain of Mycobacterium tuberculosis among Spanish HIV-infected inmates. journals.lww.com

6. Shin S, Modongo C, … RN-TJ of, et al. Advanced Immune Suppression is Associated With Increased Prevalence of Mixed-Strain Mycobacterium tuberculosis Infections Among Persons at High Risk for. academic.oup.com

7. Ssengooba W, Cobelens F, Nakiyingi L, et al. High genotypic discordance of concurrent Mycobacterium tuberculosis isolates from sputum and blood of HIV-infected individuals. ncbi.nlm.nih.gov

8. Theisen A, Reichel C, Rüsch-Gerdes S, et al. Mixed-strain infection with a drug-sensitive and multidrug-resistant strain of Mycobacterium tuberculosis. Lancet (London, England) 1995; 345:1512

9. van Rie A, Victor TC, Richardson M, et al. Reinfection and Mixed Infection Cause Changing

Mycobacterium tuberculosis Drug-Resistance Patterns. Am. J. Respir. Crit. Care Med. 2005; 172:636–642

10. Köser CU, Bryant JM, Becq J, et al. Whole-genome sequencing for rapid susceptibility testing of M. tuberculosis. N. Engl. J. Med. 2013; 369:290–292

11. Shin SS, Modongo C, Baik Y, et al. Mixed Mycobacterium tuberculosis-Strain Infections Are Associated With Poor Treatment Outcomes Among Patients With Newly Diagnosed Tuberculosis, Independent of Pretreatment Heteroresistance. J. Infect. Dis. 2018;

12. Zetola NM, Shin SS, Tumedi KA, et al. Mixed Mycobacterium tuberculosis complex infections and false-negative results for rifampin resistance by GeneXpert MTB/RIF are associated with poor clinical outcomes. J. Clin. Microbiol. 2014; 52:2422–9

13. Cohen T, Chindelevitch L, Misra R, et al. Within-host heterogeneity of mycobacterium tuberculosis infection is associated with poor early treatment response: A prospective cohort study. J. Infect. Dis. 2016; 213:1796–1799

14. Zetola NM, Modongo C, Moonan PK, et al. Clinical outcomes among persons with pulmonary tuberculosis caused by Mycobacterium tuberculosis isolates with phenotypic heterogeneity in results of drug-susceptibility tests. J. Infect. Dis. 2014; 209:1754–1763

15. Liu Q, Via LE, Luo T, et al. Within patient microevolution of Mycobacterium tuberculosis correlates with heterogeneous responses to treatment. Sci. Rep. 2015; 5:

16. Dai S, Long Y. Genotyping analysis using an rflp assay. Methods Mol. Biol. 2015; 1245:91–99

17. Schürch AC, van Soolingen D. DNA fingerprinting of Mycobacterium tuberculosis: From phage typing to whole-genome sequencing. Infect. Genet. Evol. 2012; 12:602–609

18. McIvor A, Koornhof H, Kana BD. Relapse, re-infection and mixed infections in tuberculosis disease. Pathog. Dis. 2017; 75:

19. Cohen KA, Abeel T, Manson McGuire A, et al. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. PLOS Med. 2015; 12:e1001880

20. Manson AL, Cohen KA, Abeel T, et al. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. Nat. Genet. 2017; 49:395–402

21. Votintseva AA, Bradley P, Pankhurst L, et al. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. J. Clin. Microbiol. 2017; 55:1285–1298

22. Pankhurst LJ, del Ojo Elias C, Votintseva AA, et al. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: A prospective study. Lancet Respir. Med. 2016; 4:49–58

23. van der Werf MJ, Ködmön C. Whole-Genome Sequencing as Tool for Investigating International Tuberculosis Outbreaks: A Systematic Review. Front. Public Heal. 2019; 7:87

24. Anyansi C, Keo A, Walker BJ, et al. QuantTB – a method to classify mixed Mycobacterium tuberculosis infections within whole genome sequencing data. BMC Genomics 2020; 21:80

25. Gan M, Liu Q, Yang C, et al. Deep whole-genome sequencing to detect mixed infection of

mycobacterium tuberculosis. PLoS One 2016; 11:1–14

26. Sobkowiak B, Glynn JR, Houben RMGJ, et al. Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data. BMC Genomics 2018; 19:613

27. Leinonen R, Sugawara H, Shumway M, et al. The sequence read archive. Nucleic Acids Res. 2011; 39:D19-21

28. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;

29. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078–2079

30. Walker BJ, Abeel T, Shea T, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One 2014; 9:e112963

31. Clark TG, Mallard K, Coll F, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. PLoS One 2013; 8:e83012

32. Crispell J, Zadoks RN, Harris SR, et al. Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand. BMC Genomics 2017; 18:180

33. Sampson SL. Mycobacterial PE/PPE Proteins at the Host-Pathogen Interface. Clin. Dev. Immunol. 2011; 2011:1–11

34. Casali N, Broda A, Harris SR, et al. Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. PLoS Med. 2016; 13:

35. Liu L, Jiang F, Chen L, et al. The impact of combined gene mutations in inhA and ahpC genes on high levels of isoniazid resistance amongst katG non-315 in multidrug-resistant tuberculosis isolates from China. Emerg. Microbes Infect. 2018; 7:

36. Ruesen C, Chaidir L, van Laarhoven A, et al. Large-scale genomic analysis shows association between homoplastic genetic variation in Mycobacterium tuberculosis genes and meningeal or pulmonary tuberculosis. BMC Genomics 2018; 19:122

37. Stucki D, Ballif M, Bodmer T, et al. Tracking a tuberculosis outbreak over 21 years: Strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. J. Infect. Dis. 2015;

38. Roosaare M, Vaher M, Kaplinski L, et al. StrainSeeker : fast identification of bacterial strains from unassembled sequencing reads using user-provided guide trees . bioRxiv 2016; 1–20

39. Byrd AL, Perez-Rogers JF, Manimaran S, et al. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC Bioinformatics 2014; 15:262

40. Ahn TH, Chai J, Pan C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics 2015; 31:170–177

41. Guerra-Assunção JA, Houben RMGJ, Crampin AC, et al. Recurrence due to Relapse or Reinfection With *Mycobacterium tuberculosis* : A Whole-Genome Sequencing Approach in a Large, Population-Based Cohort With a High HIV Infection Prevalence and Active Follow-up. J. Infect. Dis. 2015; 211:1154–1163

42. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 2009; 26:1641–50

43. Wiens KE, Woyczynski LP, Ledesma JR, et al. Global variation in bacterial strains that cause tuberculosis disease: A systematic review and meta-analysis. BMC Med. 2018; 16:196

44. Tientcheu LD, Haks MC, Agbla SC, et al. Host Immune Responses Differ between M. africanum- and M. tuberculosis-Infected Patients following Standard Anti-tuberculosis Treatment. PLoS Negl. Trop. Dis. 2016;

45. De Jong BC, Adetifa I, Walther B, et al. Differences between tuberculosis cases infected with Mycobacterium africanum, West African type 2, relative to Euro-American Mycobacterium tuberculosis: An update. FEMS Immunol. Med. Microbiol. 2010;

46. de Jong BC, Hill PC, Aiken A, et al. Progression to Active Tuberculosis, but Not Transmission, Varies by Mycobacterium tuberculosis Lineage in The Gambia . J. Infect. Dis. 2008;

47. Tarashi S, Fateh A, Mirsaeidi M, et al. Mixed infections in tuberculosis: The missing part in a puzzle. Tuberculosis 2017; 107:168–174

48. Fiebig L, Kohl TA, Popovici O, et al. A joint cross-border investigation of a cluster of multidrug-resistant tuberculosis in Austria, Romania and Germany in 2014 using classic, genotyping and whole genome sequencing methods: Lessons learnt. Eurosurveillance 2017; 22:

49. Merker M, Blin C, Mona S, et al. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nat. Genet. 2015; 47:242–249

50. Ribeiro SCM, Gomes LL, Amaral EP, et al. Mycobacterium tuberculosis strains of the modern sublineage of the Beijing family are more likely to display increased virulence than strains of the ancient sublineage. J. Clin. Microbiol. 2014; 52:2615–24

51. Rutaihwa LK, Menardo F, Stucki D, et al. Multiple introductions of Mycobacterium tuberculosis Lineage 2-Beijing into Africa over centuries. Front. Ecol. Evol. 2019;

52. Rinder H, Mieskes KT, Löscher T. Heteroresistance in Mycobacterium tuberculosis. Int. J. Tuberc. Lung Dis. 2001;

53. Eilertson B, Maruri F, Blackman A, et al. High proportion of heteroresistance in gyra and gyrb in fluoroquinolone-resistant mycobacterium tuberculosis clinical isolates. Antimicrob. Agents Chemother. 2014;

54. Folkvardsen DB, Thomsen VO, Rigouts L, et al. Rifampin heteroresistance in Mycobacterium tuberculosis Cultures as Detected by phenotypic and genotypic drug susceptibility test methods. J. Clin. Microbiol. 2013;

55. . WHO | WHO treatment guidelines for multidrug- and rifampicin-resistant tuberculosis, 2018 update. WHO 2019;

56. Kamakoli MK, Sadegh HR, Farmanfarmaei G, et al. Evaluation of the impact of polyclonal infection and heteroresistance on treatment of tuberculosis patients. Sci. Rep. 2017;

4 Global antibiotic resistance acquisition patterns of *Mycobacterium tuberculosis*

4.1 Abstract

The emergence and continued growth of multi-drug (MDR) and extensively drug resistant (XDR) tuberculosis has been a growing concern in the past decade due to their ability to impact treatment of this infectious disease. As the number of cases of antibiotic resistant tuberculosis continues to rise, whole-genome sequencing based research has been widely used to provide insights into the genetic mechanisms driving the evolution of antibiotic resistance.

In this study we used a collection of more than 40,000 Mycobacterium tuberculosis isolates to analyze the acquisition order of antibiotic resistance mutations. We constructed a phylogenetic tree from our globally diverse set and used a maximum likelihood based method to reconstruct the evolutionary path of mutations throughout the phylogenetic tree. Our data set has a large number of MDR and XDR isolates. We show that in the majority of cases mutations conferring resistance to isoniazid, precedes that of rifampicin. Due to the size of our dataset, we were able to determine orders of resistance for less frequent drug mutations as well. Our research corroborates previous research findings that drug resistance evolves in an expected predictable order on a global scale and extends the ordering beyond first-line anti-tubercular drugs.

4.2 Introduction

In 2019, tuberculosis (TB) caused a total of 1.4 million deaths and infected 10 million new people, making it one of the top 10 most deadly infectious agents [1]. The rapid emergence of antibiotic resistant TB poses a major hurdle in effectively treating the disease. Incidence of multidrug-resistant TB (MDR-TB), a form of TB which has shown to be resistant to both isoniazid and rifampicin, two of the most commonly used drugs, has increased by 10% from 2018 to 2019. Only 57% of patients with MDR-TB receive successful treatment [1]. An even more daunting threat is the rise of extensively drug resistant TB (XDR-TB), defined by the WHO in 2019 to be a form of TB resistant to isoniazid and rifampicin in addition to two other potent anti TB drugs. More than 10,000 cases of XDR-TB was identified in 2017 across 77 countries [1]. Early diagnosis of drug resistance in TB patients can influence proper treatment and reduce the spread of antibiotic resistant TB throughout the globe.

Due to the increasing use of whole genome sequencing (WGS), an abundance of new information and insights about the *Mycobacterium tuberculosis* genome and its impact on human infection have been uncovered[2,3]. Using WGS, researchers have identified specific variations of genes within the MTB genome that lead to phenotypic antibiotic resistance[4–6]. Numerous databases have been developed to provide researchers with data on the mutations leading to antibiotic resistance[7–10]. This has led to researchers relying on WGS to study antibiotic resistance transmission within populations, as the detection of antibiotic resistant cases of TB from WGS

provides advantages over the traditional susceptibility testing methods which can be time consuming or limited in their sensitivity for lower prevalence resistance [11].

The most widely used antibiotic resistance diagnostic, GeneXpert, detects the presence of rifampicin resistance. However, previous research looking into trends in the order of antibiotic resistance acquisition in TB have found that resistance to isoniazid overwhelmingly precedes all other drug resistances including rifampicin[12–14]. Therefore, by the time GeneXpert detects rifampicin resistance, a patient most likely has already developed resistance to isoniazid or other drugs. Detecting pre-MDR or pre-XDR cases of MTB would enable pro-active response and treatment to TB.

Previous research has been conducted to examine the development of MDR and XDR TB. Findings from these studies indicate certain drugs or mutations are more likely to provide an earlier gain of fitness for the bacteria. These studies are however, limited in the size of their datasets thereby limiting their ability to resolve the order of drug resistance acquisition for less frequently observed drug resistances. It is important to know which drug mutations are more likely to occur earlier than others across all drugs, in order to better understand which mutations provide a fitness advantage for TB, obtain a clearer picture into the development of MDR-TB and XDR-TB, as well as provide a better view into possible treatment options based on the mutation resistance profile.

In this study, we used a large dataset of 40,088 MTB strains to investigate the general resistance acquisition order observed in MTB. Using a maximum likelihood based ancestral reconstruction model, we estimated and compared events of resistance acquisition within a phylogenetic context. We resolved the likely order of drug resistance mutations for MTB across 54 different countries and 7 different antibiotic resistance drugs. We also identified harbinger antibiotic resistance mutations which commonly co-occur with the other, as well as mutations which commonly co-evolve with each other at similar positions within a phylogenetic tree.

4.3 Methods

4.3.1 Data collection

53,025 WGS short-read datasets were downloaded from NCBI's sequencing read archive (SRA) [15] on December 12, 2018 using the taxonomic ID for *M. tuberculosis:* txid1494075. Samples were filtered and quality controlled as previously described. Briefly, Illumina data only, read length >= 50 bp, coverage >= 20x. Samples with possible mixed infections were excluded after identification with QuantTB [16]. After quality control filters and removing mixed samples, the final working dataset contained 40,088 samples.

Geographic region of isolation for 69% of all SRA samples could be determined based on their BioSample metadata identifier (37% of isolates) or from their corresponding publication (an additional 33% of isolates), which we examined for studies with at least 100 isolates.

4.3.2 Variant calling

Variant calling was performed as described before. Briefly, SRA files were extracted using fastqdump (version 2.90) [15], then aligned against the *M. tuberculosis* H37Rv reference genome (GenBank: CP003248.2) with BWA-MEM (Version: 0.7.17-r1188) [17] using default settings, then index-sorted with samtools (Version: 1.6 , using htslib 1.6) [18] Finally variants were called using Pilon (version 1.22, default settings, fixes set to none) [19].

4.3.3 Drug resistance prediction

Genotypic antibiotic resistance to seven drugs was identified using QuantTB based on a curated set of SNPs [12]. Only positions conferring a homoresistant (resistance-conferring allele at a frequency of > 90%), or heterozygous resistant (two different resistance-conferring alleles present at a summed frequency of >90%) were considered in this analysis. 3,451 positions conferring heteroresistance to a drug (positions containing alleles for both resistant and susceptible genotypes where at least one resistance-conferring allele was present at a frequency between 10-90%) were excluded as this is indicative of within host evolution, which could influence the accuracy of our results. A sample was considered susceptible to a drug if it contained no resistance-conferring alleles at frequency of at least 10%.

4.3.4 Ordering of drug resistance

A bootstrapped phylogenetic tree was constructed for the dataset based on an alignment of genomic positions containing SNPs identified by Pilon. Insertions, deletions, bases with low quality (Phred score less than 11) and bases within PE/PPE regions (Supplementary Table 1) were not considered. FastTree [20] was used to generate a tree from the remaining SNPs with default settings, and used to calculate local support values for each node.

The function *ace* from the *ape* package [21] was used to reconstruct the gains and losses of mutations at each node of the midpoint rooted phylogenetic tree. As there is more evolutionary pressure for a mutation to be gained than lost, *ace* was run using a transition matrix that applied a 2x penalty for a loss of a resistance mutation than a gain of a resistance mutation. A cost of 2x was chosen as the community consensus is that gain of a resistance mutation is more likely than the loss of such a mutation due to its fitness benefit under antibiotic pressure. Presence of a particular mutation at a node was only called if the likelihood for that mutation was above 70%. 70% was chosen to ensure that mutation predictions were trustworthy.

To estimate the order of drug resistance of different drugs, we captured the predicted states from *ace* for each node in the sequence of nodes from the root of the tree to each tip containing a sample with had an antibiotic resistance conferring mutation.

4.4 Results

4.4.1 Global distribution of antibiotic resistance *M. tuberculosis* strains

In order to determine patterns of MTB antibiotic resistance acquisition, we compiled a dataset of 40,088 MTB whole genome sequencing datasets. Our dataset is globally distributed, covering 106 different countries and territories, with the majority of isolates originating from the UK (6,922), South Africa (3,554), and Malawi (1820). All lineages (1-7) are represented in our dataset. Unsurprisingly, Lineages 4 and 2 were the most abundant in our dataset as these lineages are the most frequently observed across the globe [22].

For each of the MTB isolates, we determined genotypic resistance to 9 TB drugs based on a list of 143 drug resistance mutations (see Methods). Mutations for antibiotic resistance were observed in 68% of the 107 countries represented in our dataset. Out of 40,088 isolates, 34% (12,909) harbored antibiotic resistance to one or more drugs, and the majority of these (9,597) harbored antibiotic resistance to more than one drug. The heatmap in Figure 4.1 B shows clearly that many strains (rows) have multiple drug resistance mutations. As expected, resistances to the typical first line antibiotics were observed in the greatest quantities, with rifampicin (27%) and isoniazid (27%) the most common, followed by streptomycin (21%) and ethambutol (18.5%) (Table 4.1). These antibiotics are the typical first line drugs used to treat MTB which leads to their high observance in our dataset. Specifically streptomycin, which is the oldest anti-tubercular drug, has been administered since 1944[23]. Resistance to pyrazinamide, another first line drug was not as highly present in our dataset. Lack of pyrazinamide resistance in our dataset is not related to its epidemiological incidence, but rather reflects our incomplete knowledge of the genetic mechanisms of its antibiotic resistance. Despite rifampicin resistance being more abundant than isoniazid resistance, the most commonly observed antibiotic resistance mutation was katG S315T, conferring resistance to isoniazid in 26% (10,385) of samples, followed by rpoB S450L, conferring resistance to rifampicin in 18% (7,392) of samples (Table 4.2).

A large proportion of multi-drug resistant (MDR) isolates, resistant to both rifampicin and isoniazid, was observed. MDR isolates accounted for nearly 50% of those containing antibiotic resistance and 16% of all isolates. Levels are in line with global WHO rates of MDR present in previously treated cases (19%). Highest rates of MDR amongst countries with at least 30 samples were observed in Argentina (92%, 170/183), Bangladesh (84%, 38/45), and Sweden (69%, 88/128). We observed higher rates of extremely drug resistant TB among MDR cases compared to global WHO rates (12% vs 6%). Countries in the Caucasus exhibited highest rates of XDR-TB: Azerbaijan (62/213, 29%), Belarus (45/155, 29%), and Georgia (46/308, 15%). Although the large proportion of MDR-TB and XDR-TB samples within our dataset doesn't necessarily reflect global trends, it does reflect the large focus on understanding resistance within TB research and the studies we have surveyed.

Table 4.1: Incidence rates of resistance mutations for each of the seven drugs analyzed.

Drug	# isolates	# Mono-resistant isolates[1]	% of resistant isolates with mono-resistance[2]	% of all isolates with this resistance[3]
rifampicin	10,929	734	7%	27%
isoniazid	10,651	968	9%	27%
streptomycin	8,463	858	10%	21%
ethambutol	7,422	65	1%	19%
common to isoniazid & ethionamide	4,059	1,026	25%	10%
ofloxacin	2,782	158	6%	7%
kanamycin	2,641	23	1%	7%
pyrazinamide	782	9	1%	2%
ethionamide	154	1	1%	<1%

[1]Number of isolates that are mono-resistant, i.e. only resistant to the corresponding drug
[2]Proportion of isolates harboring resistance to the corresponding drug, that are mono-resistant for the drug
[3]Proportion of all isolates in the dataset that harbor genotypic resistance to the corresponding drug

Table 4.2 Antibiotic resistance mutation counts for the ten most abundant mutations

Position: Allele	Mutation	Drug	Gene	Codon change	# of samples	% of all isolates
2155175:G	S315T	Isoniazid	*katG*	Ser-315-Thr	10385	26%
761158:T	S450L	rifampicin	*rpoB*	Ser-450-Leu	7392	18%
781690:G	K43R	streptomycin	*rpsL*	Lys-43-Arg	5238	13%
1673431:T		Common to isoniazid and ethionamide	*inhA_p*		3348	8%
4247607:G	M306V	ethambutol	*embB*	Met-306-Val	2763	7%
1473252:G		kanamycin	*rrs*		1742	4%
4247609:A	M306IA	ethambutol	*embB*	Met-306-Ile	1458	4%
781825:G	K88R	streptomycin	*rpsL*	Lys-88-Arg	1193	3%
7582:G	D94G	Ofloxacin	*gyrA*	Asp-94-Gly	973	2%
1472365:C		streptomycin	*rrs*		786	2%

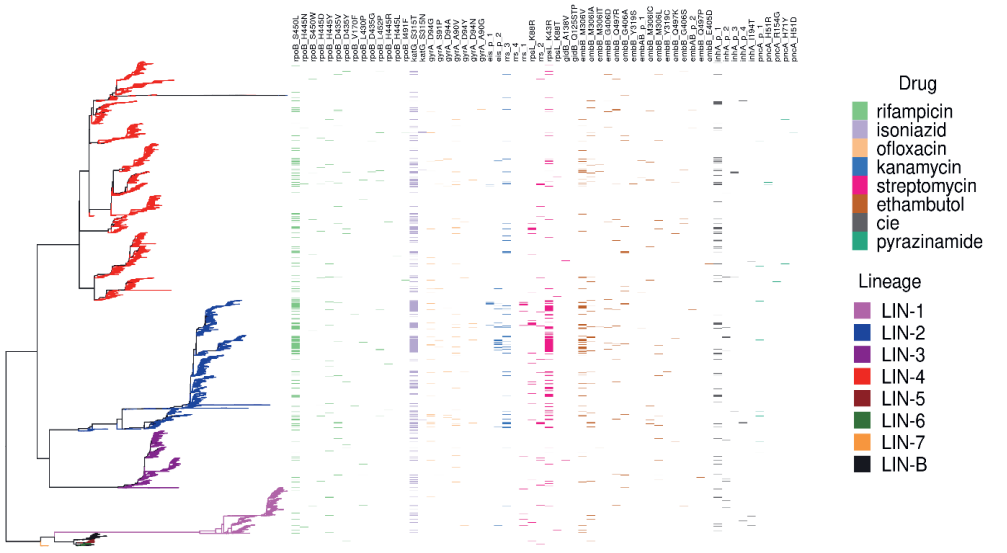


Figure 4.1 Phylogeny of 40,088 MTB isolates (left) with a heatmap corresponding to the antibiotic resistance mutations present in the tips of the phylogeny is shown (right). The phylogeny was made from an alignment of concatenated SNP sequences input into FastTree and midpoint rooted. Mutations for each of the drugs are shown for every isolate in the phylogeny, colored by drug the mutation is conferring resistance to. CIE refers to mutations that are common to both isoniazid and ethionamide. Tip nodes in phylogeny are colored according to lineage of the corresponding isolate.

4.4.2 Evolution of drug resistance

In order to quantify the order of drug resistance mutation acquisitions, we applied a maximum likelihood ancestral state reconstruction method to the phylogeny of the 40,088 isolates. At each node of the phylogeny, gain and/or loss for each antibiotic resistance mutation was predicted for each of the 52 mutations present in greater than 50 isolates (see Methods). We observed that independent resistance acquisition events, or the number of unique nodes in which a gain of mutation occurred, was not specific to a particular country or region of the phylogeny, suggesting drug resistant mutation acquisition occurred in a similar fashion globally. This finding has previously been noted in prior research [12]. The majority of independent resistant acquisition events were at the tip nodes (60%), i.e. the resistant isolate. The remaining 40% indicate possible transmission of antibiotic resistance or outbreaks of drug resistant isolates.

4.4.3 Drug resistance acquisition order

We analyzed the order of antibiotic resistance acquisition for every resistant tip on the phylogeny to determine the ordering of acquisition events To resolve acquisition chains, i.e. the evolutionary sequence of resistance mutation acquisitions for a specific isolate, we examined the path containing resistance mutations gained or lost from each isolate tip to the node in the phylogeny where the first resistance was acquired. For example, if a certain tip harbored resistance to ethionamide, its acquisition chain would consist of all ancestral nodes of that tip that contain the

specific mutation conferring the observed resistance. Examining these resistant acquisition chains not only gives insight into the evolutionary trajectory of these mutations, but also enables the identification of clustering of clonal expansions of different isolates.

For this analysis, we excluded isolate chains where resistance was acquired on the tip node as there is no sequential information to be gained from these isolates. There were 6,316 unique acquisition chains present within our phylogeny. Table 4.3 shows the ordering of the ten most common chains that arose independently throughout the tree, where more than one resistance was gained. The *katG* S315T mutation has previously been said to develop prior to resistance to any other drug. *KatG* S315T followed by *rpoB* S450L occurred in 752 strains, making it the most frequently observed resistance acquisition chain. This acquisition chain occurred 429 times independently across our dataset, indicating the high propensity for *katG* S315T to precede *rpoB* S450L for MDR resistance, as observed in past research [12,13]. Indeed, more than half (57%) of the acquisition chains with the greatest amount of evolutionary independent arisals in our dataset (# unique events > 50) start with the *katG* S315T mutation. This is followed by the *rpsL* K43R mutation (17%) and the *rpoB* S450L mutation (15%).

On the other hand, we observed acquisition chains found frequently in our dataset (> 50 strains) that arose from a few independent acquisition events. This was the case for the acquisition chain *inhA promoter* -> *katG* S315T -> *rpoB* S450L/ *rpsL* K88R, which arose 12 times independently in 83 strains. Usually, such findings can partly be explained by local transmission events or clonal expansion of certain strains, but the samples present within both these sets originate from multiple studies (15) and geographic locations (8). It is likely that these frequently observed acquisition chains are part of a large transmission event of a highly resistant and widely circulating strain.

Table 4.3 Ten most frequent resistance acquisition chains. Each row shows the path of the resistance mutation gains in the numbered columns. The number of strains indicates how many strains this particular acquisition chain was noted in. The number of unique events indicates the number of distinct (non-overlapping) nodes responsible for the chain, i.e. the number of monophyletic groups containing the chain.

First	Second	Third	# Strains	Unique events
katG S315T	*rpoB* S450L		752	296
katG S315T	*embB* M306V		249	111
rpsL K43R	*rpoB* S450L		274	109
katG S315T	*rpsL* K43R		237	124
katG S315T	*embB* M306IA		223	118
rpsL K43R	*katG* S315T		223	92
katG S315T	*rpoB* D435V		134	87
rpoB S450L	*katG* S315T		217	84
katG S315T	*rrs* 1		146	52
katG S315T	*rpoB* S450L	*eis* promoter 2	110	37

In order to determine the relative ordering between antibiotic resistance mutation pairs we examined the acquisition chains to identify the proportion of times a mutation evolved prior to every other mutation (Figure 4.2). Similar to previous research, we found that isoniazid resistance (mutations in *katG* S315T) overwhelming precedes all other antibiotic resistance mutations.

Interestingly, the *rpsL* K43R mutation, conferring resistance to streptomycin, precedes resistance to other antibiotic mutations just as frequently as *katG* S315T mutation. *KatG* S315T however, occurs before *rpsL* K43R most of the time, although the difference is not as significant as it is compared to the other mutations (60% of the time both mutations are in a unique sequence, *katG* S315T precedes *rpsL* K43R). Both *katG S315T* and *rpsL* K43R precede overwhelmingly all other types of rifampicin resistance, this is especially true for *katG* S315T which precedes rifampicin mutations on average 97% of cases, while *rpsL* K43R precedes rifampicin in 91% of cases. Extrapolating to all streptomycin resistances shows the same pattern, where on average resistance to streptomycin develops prior to that of rifampicin 83% of the time. The mutation in *inhA promoter,* which confers resistance to both isoniazid and ethioniamide, also preceded rifampicin resistance in the majority of cases (80% of the time), confirming the previous observation that MDR resistance typically starts with isoniazid resistance which is later followed by rifampicin resistance.

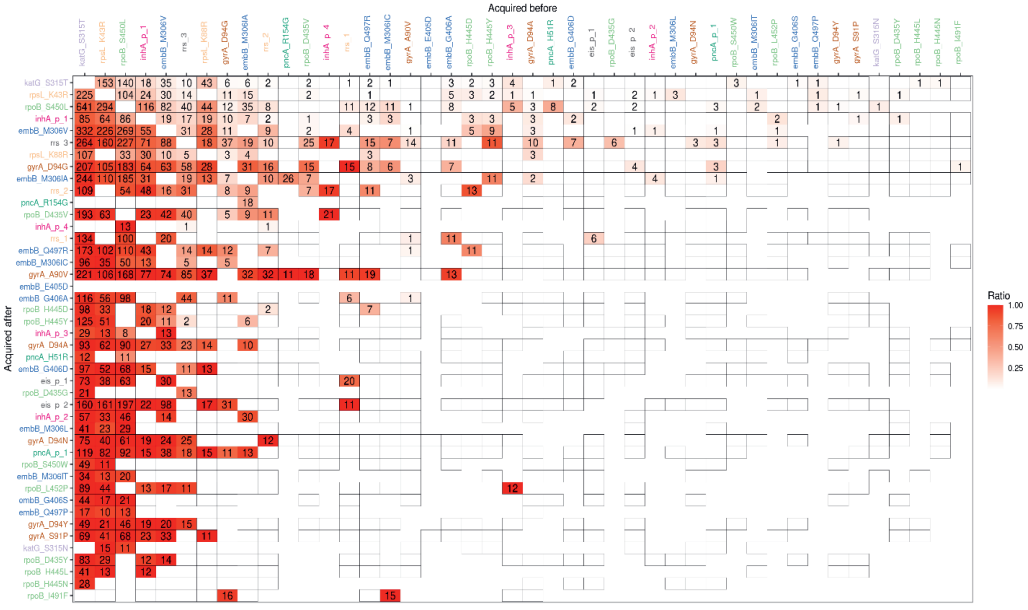


Figure 4.2 : Pairwise arisals of antibiotic resistance mutations present in our dataset. Values correspond to the number of times mutation in the X axis is acquired before the corresponding mutation in the Y axis. Only pairs with occurring in either direction at least 10 times are shown. Colored are the proportion of times this occurs relative to the total number of unique pairs containing the two mutations. Cells are shaded relative to the intensity of proportion (darker red denoting proportions closer to 1, while lighter reds indicate less frequent proportions closer to 0) White cells indicate the mutation on the x axis never is acquired before the mutation in the y axis, while empty cells indicate the pair never occurs in the same sequence. Resistance mutations are colored based on the drug they confer resistance to identical to Figure 4.1.

The size of our dataset also allows sufficient resolution to elucidate acquisition order of mutations lesser abundant in the population. Our data suggests certain mutations that overwhelmingly evolve after any another mutation, namely mutations conferring resistance to ofloxacin and pyrazinamide. Both the mutations in *gyrA* A90V and the *pncA* promoter occurred last in all of the

mutation pairs they occurred in (14 and 9 respectively). Overall, mutations conferring resistance to pyrazinamide and ofloxacin evolved last on average 90% and 87% of the time respectively. As resistance to ofloxacin is one of the drugs indicative of XDR resistance, it is likely that MTB isolates develops resistance to ofloxacin more slowly compared to other drugs. Mutations leading to another second line drug indicative of XDR, kanamycin, evolves on average prior ofloxacin, suggesting that XDR resistance begins with isoniazid, followed by rifampicin, then kanamycin and lastly ofloxacin. Based on our data, the ordering of the 8 drugs surveyed in our dataset is as depicted in Figure 4.3.

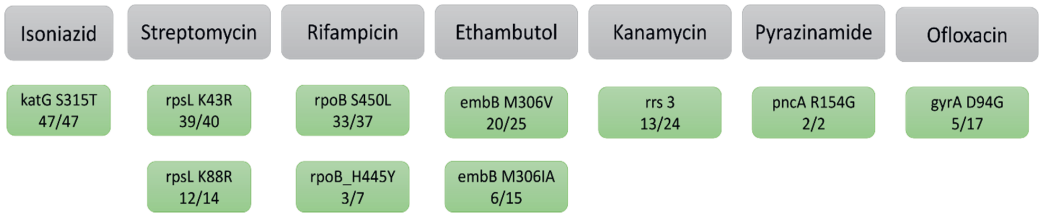


Figure 4.3 Phylogeny based acquisition ordering of antibiotic resistance mutations in MTB isolates. Starting from first acquired (left; isoniazid) to last acquired (right; ofloxacin). Colored in green are the specific mutations, leading to resistance of its drug (colored in grey in the first row), that are most frequently observed first within a pair of antibiotic resistance mutations. The proportion of times it is observed first in a pair of antibiotic resistant mutations (shown in a pairwise manner in Figure 5.1Figure 4.2) is reported as a fraction. E.g. the *katG* S315T mutation occurs in an acquisition chains with 47 other mutations. This mutation evolves before each of these other mutations, leading to a ratio of 47/47.

4.5 Conclusions

Using 40,088 MTB isolates we examined trends in genotypic antibiotic resistance covering all major lineages and geographic regions. This diverse collection of isolates represents the largest study of antibiotic resistance acquisition in MTB to date. The data set has an overrepresentation of antibiotic resistant isolates, compared to WHO estimates, reflecting the global interest in resolving the genetic origins of drug resistance.

Different antibiotic resistance mutations have varying levels of fitness cost for MTB, thereby suggesting that antibiotic resistance is developed in a particular order. Further examination of antibiotic resistance mutation acquisition leading to MDR or XDR is especially helpful for researchers and healthcare professionals in order to refocus their treatment/research goals strategies such that treatment is in line with the propensity of a patient to become antibiotic resistant and to better understand why certain antibiotic resistances are more likely to precede others. Previous research looking into the acquisition of MTB antibiotic resistance mutations have found resistance to isoniazid generally precedes that of rifampicin [13,14]. We corroborate and extend this finding to by resolving the order in which antibiotic resistance mutations develop in TB: (1) isoniazid, (2) streptomycin, (3) rifampicin, (4) ethambutol, (5) kanamycin, (6) pyrazinamide and (7) ofloxacin.

One explanation for the development of isoniazid resistance before all other resistances is that the mutation rate of the gene that contains the isoniazid resistance mutation is generally higher than other genes containing resistance conferring mutations [24]. However this has been shown to be not an adequate explanation as in vitro determined rates of isoniazid mutations does not reflect clinical rates [14,24,25]. Another explanation for early isoniazid acquisition is based on the fact that different drugs have different fitness costs for the bacteria. The katG S315T mutation is not only the most common path to isoniazid resistance but it also carries a relatively low fitness burden for MTB likely contributing to the early arisal of this resistance compared to other drugs[13,26]. Because of the size of our dataset, we were also able to discover that resistance to streptomycin likely occurs before rifampicin, a finding not yet explored in previous research. This finding is quite interesting as streptomycin is not a first line drug and not used as frequently as rifampicin. The arisal of streptomycin resistance before other drugs, is most likely an ancestral artefact of TB due to the long exposure of streptomycin to MTB as streptomycin was the first successful drug administered to treat TB.

By examining chains of resistance acquisition, we identified clusters of samples with resistances that descended from the same acquisition event. As these samples originated from varying biological studies and geographical locations, incidence of such clusters point to the transmission of antibiotic resistance strains to different regions. This phenomena was observed in prior studies which found 17 clusters of MDR-resistant strains transmitted across geographical borders [27]. Through further examination of these acquisition chains we confirmed previous research that resistance to isoniazid develops prior to all other mutations. This observation has held true across different geographic regions and lineages in previous studies of MTB resistance acquisition[12–14].

One limitation in our study is due to the inadequate knowledge of antibiotic resistance mutations for TB. There is an active field of study to determine which mutations lead to antibiotic resistance in TB [4,5,28]. In our study we used one of the most recent manually curated lists of mutations to make predictions for antibiotic resistance [16]. Our list has flaws especially due to the limited knowledge of pyrazinamide resistance [29]. It is likely that we underestimated the amount of antibiotic resistance in our dataset. Future studies with more complete antibiotic resistance mutation lists would add upon our study in clarifying the order of drugs with lesser known genetic basis of resistance.

In conclusion, our results shed light on the emergence of antibiotic resistance on a global scale. We corroborate previous research that found that drug resistance evolves in a particular order within TB. We observed a few of the most common drug acquisition chains which have held valid across our dataset, particularly with isoniazid resistance preceding rifampicin and streptomycin resistance. Further research into the biological mechanisms of antibiotic resistance could help clarify the possible interactive effects different drugs have on each other and lead to better treatment protocols.

4.6 References

1. World Health Organisation. Global Tuberculosis Report 2019. 2019;

2. Miyahara R, Smittipat N, Juthayothin T, et al. Risk factors associated with large clusters of tuberculosis patients determined by whole-genome sequencing in a high-tuberculosis-burden country. Tuberculosis 2020;

3. Katale BZ, Mbelele PM, Lema NA, et al. Whole genome sequencing of Mycobacterium tuberculosis isolates and clinical outcomes of patients treated for multidrug-resistant tuberculosis in Tanzania. BMC Genomics 2020;

4. Walker TM, Kohl TA, Omar S V, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect. Dis. 2015; 15:1193–1202

5. Farhat MR, Sultana R, Iartchouk O, et al. Genetic determinants of drug resistance in mycobacterium tuberculosis and their diagnostic value. Am. J. Respir. Crit. Care Med. 2016;

6. Zignol M, Cabibbe AM, Dean AS, et al. Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. Lancet Infect. Dis. 2018;

7. Ghosh A, Saran N, Saha S. Survey of drug resistance associated gene mutations in Mycobacterium tuberculosis, ESKAPE and other bacterial species. Sci. Rep. 2020;

8. Miotto P, Tessema B, Tagliani E, et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. Eur. Respir. J. 2017;

9. Sandgren A, Strong M, Muthukrishnan P, et al. Tuberculosis drug resistance mutation database. PLoS Med. 2009;

10. Flandrois JP, Lina G, Dumitrescu O. MUBII-TB-DB: A database of mutations associated with antibiotic resistance in Mycobacterium tuberculosis. BMC Bioinformatics 2014;

11. Kim SJ. Drug-susceptibility testing in tuberculosis: Methods and reliability of results. Eur. Respir. J. 2005; 25:564–569

12. Manson AL, Cohen KA, Abeel T, et al. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. Nat. Genet. 2017; 49:395–402

13. Cohen KA, Abeel T, Manson McGuire A, et al. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. PLOS Med. 2015; 12:e1001880

14. Ektefaie Y, Dixit A, Freschi L, et al. Tuberculosis resistance acquisition in space and time: an analysis of globally diverse M. tuberculosis whole genome sequences. bioRxiv 2019; 837096

15. Leinonen R, Sugawara H, Shumway M, et al. The sequence read archive. Nucleic Acids Res. 2011; 39:D19-21

16. Anyansi C, Keo A, Walker BJ, et al. QuantTB – a method to classify mixed Mycobacterium tuberculosis infections within whole genome sequencing data. BMC Genomics 2020; 21:80

17. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;

18. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078–2079

19. Walker BJ, Abeel T, Shea T, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One 2014; 9:e112963

20. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 2009; 26:1641–50

21. Paradis E, Schliep K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2019; 35:526–528

5 Reconstruction of PE/PPE genes in high-quality M. tuberculosis genome assemblies reveals lineage specific distribution patterns

5.1 Abstract

The *Mycobacterium tuberculosis* genome includes the PE and PPE gene families, whose functions largely remain unclear despite representing 10% of the genome. These genes are highly repetitive and GC-rich making them difficult to sequence and assemble using short read sequencing technologies. Previous studies have begun to look into the important role these genes play in bacterial virulence and cellular host immune response. However, these studies have been constrained by surveying only a subset of the more than 160 PE/PPE genes in lineages of *M. tuberculosis* with limited diversity. Here we used long read technology to sequence and assemble 18 genomes covering seven human-adapted *M. tuberculosis* lineages (1-7) from across the globe. These were supplemented with three genome assemblies: *Mycobacterium canetti,* a lineage 8 strain, and a high-quality reference of H37Rv. With these 21 genomes spanning all known *M. tuberculosis* lineages we conducted a comparative study of variation within and between all PE/PPE genes. Multiple sequence alignments of each gene revealed genomic variations that are specific to lineages and identified genes that are highly conserved within the two families. We report four novel PE/PPE genes that are missing from the H37Rv reference annotation. Moreover, we provide high quality reference assemblies as a resource for future studies of the *M. tuberculosis* genome. With this study, we provide the genetic basis for more in-depth mechanistic studies into the complex organization of PE/PPE family members and their role in virulence, pathogenesis and their use as vaccine targets.

5.2 Introduction

Mycobacterium tuberculosis, the causative bacteria for the disease Tuberculosis (TB), is one of the most deadly infectious agents ever known to mankind - responsible for the deaths of millions each year and affecting the lives of nearly 2 billion people across the globe [1]. Organisms within the *M. tuberculosis complex* (MTB) are characterized by their relatively low mutation rate, resulting in a clonal population structure with low levels of sequence diversity [2]. Strains of *M. tuberculosis* have been categorized into six major lineages based on their genotypic diversity: three ‘modern’ lineages (lineage 2 East-Asian, lineage 3 East-African, and lineage 4 Euro-American), and three ‘ancient’ ancestral lineages (lineage 1 Indo-Oceanic, lineage 5 Africanum Subtype 1, and lineage 6 Africanum Subtype 2) [3,4]. Recently, two additional lineages have been identified, including one lineage intermediate between the modern and ancestral lineages (lineage 7) and one lineage more distantly related to *Mycobacterium canettii* (lineage 8) [5]. Genetic diversity across these lineages has been associated with diversity in phenotypic traits with modern lineages exhibiting increased disease virulence and severity compared to other lineages [6].
Nearly 10% of the *M. tuberculosis* genome consists of two large gene families, called the PE and PPE gene families, whose naming is based on the presence of a Pro-Glu (PE) or Pro-Pro-Glu

(PPE) motif within the first 10 amino acids of its N-terminal domain [7]. The 101 genes within the PE superfamily are further subdivided into two subfamilies: the largest, PE-PGRS, contains 64 genes and is characterized by a polymorphic C-terminal domain with multiple repeats of Gly-Gly-Ala or Gly-Gly-Asn motifs, while the smallest, PE, contains only the PE domain [8].

Numerous studies have explored the role of PE and PPE proteins within *M. tuberculosis* pointing to associations between these proteins with disease pathogenicity and progression. Loss of the PPE 25-PE 19 gene cluster has been associated with reduced virulence [9], while increased cytotoxicity and inflammation was observed following expression of some PE-PGRS proteins (PE-PGRS33, PE-PGRS16, PE-PGRS26, PE-PGRS30, PE-PGRS11, and PE-PGRS17) [8]. Further, it has been speculated that genetic variation across different subsets of PE and PPE genes allows *M. tuberculosis* to either evade or modulate the host immune system [8] as some PE and PPE proteins are surface exposed antigens, enabling direct interaction with the host. PPE37 has been shown to hinder macrophage driven inflammatory response [10] and the co-expression of PE9/10 has been shown to induce macrophage apoptosis [11].

Despite accounting for nearly 10% of *M. tuberculosis* coding potential [8] PE and PPE genes are regularly excluded from whole genome sequencing analysis studies as their highly repetitive nature and abundance of GC-rich regions make them typically poorly covered by short read sequencing technologies [12]. Because genetic variations within these genes are assumed to contribute to differences in virulence [13], there is a need to properly characterize PE and PPE sequence variation within and across *M. tuberculosis* lineages. Long read sequencing technologies offers read lengths that can resolve longer highly repetitive regions that are difficult to sequence with short read technologies. Previous studies investigating variation within these gene families have yielded conflicting results, limited in either the breadth of lineages they cover, the number of genes considered, or the relatedness of isolates analyzed [7,14–17].

To better understand variation within PE and PPE genes we used the Oxford Nanopore Technologies MinION platform to sequence and assemble 18 isolates spanning seven known *M. tuberculosis* lineages sourced from a MTB clinical strain reference set [18]. Our analysis was supplemented with the high quality H37Rv assembly [19], *M. canetti*[20], and lineage 8 assembly [5]. This represents the first analysis of all 168 fully assembled PE and PPE genes from eight human adapted lineages of *M. tuberculosis*. Additionally, we provide 18 fully assembled high quality MTB genomes as a resource for the scientific community. With this study, we provide a platform for future research efforts into the functional and pathogenic properties of the *M. tuberculosis* genome.

5.3 Materials and Methods

5.3.1 Sample information

18 *M. tuberculosis* isolates covering various global regions and all seven human adapted lineages were selected for analysis based on a previously published clinical strain reference set [18]. Characteristics of these isolates are provided in Table 5.1. Three additional genomes – H37Rv (NCBI accession number AL123456.3), *M_canettii* (NCBI accession number HE572590.1), and

a lineage 8 strain (NCBI accession number ASM1292376V1) were additionally included in our analysis.

5.3.2 DNA extraction and sequencing

DNA was extracted using the Genomic DNA Buffer Set (Qiagen Inc, Germantown, Maryland, USA) as undertaken previously for long reading sequencing of the L8 strain[5].

Approximately 1 µg of DNA of each sample was used for library preparation using the Native Barcoding kit (EXP-NBD104 and EXP-NBD114), and the Ligation Sequencing kit (SQK-LSK109) following manufacturer instructions. The prepared library was loaded into an R9.4 SpotON Flow Cell and sequenced on a MinION device.

Table 5.1 Details of the samples used for comparative analysis, including 18 isolates sequenced for this study and three supplementary genomes: H37Rv (NCBI accession number AL123456.3), an *M. canettii* strain (NCBI accession number HE572590.1), and a lineage 8 strain (NCBI accession number ASM1292376V1) [43]

Sample Name	Strain Name	Country of origin	Lineage	GenBank	SRA (long reads and short read)
ITM-2018-00082	N0157	Philippines	L1	CP069078	SRR12801738
ITM-2018-00083	N0072	India	L1	CP069077	SRR12801737
ITM-2018-00087	N0031	China	L2	CP069076	SRR12801728
ITM-2018-00088	N0155	China	L2		SRR12801727
ITM-2018-00089	N0004	India	L3	CP069075	SRR12801726
ITM-2018-00090	N1274	Afghanistan	L3	CP069074	SRR12801725
ITM-2018-00091	N0054	Ethiopia	L3	CP069073	SRR12801724
ITM-2018-00092	N1216	Ghana	L4	CP069072	SRR12801723
ITM-2018-00093	N0136	USA	L4	CP069071	SRR12801722
ITM-2018-00094	N1283	Germany	L4	CP069070	SRR12801721
ITM-2018-00095	N1176	Ghana	L5	CP069069	SRR12801736
ITM-2018-00096	N1272	Ghana	L5	CP069068	SRR12801735
ITM-2018-00097	N1268	Sierra Leone	L5	CP069067	SRR12801734
ITM-2018-00098	N1201	Ghana	L6	CP069066	SRR12801733
ITM-2018-00099	N0091	Gambia	L6	CP069065	SRR12801732
ITM-2018-00100	N1202	Ghana	L6	CP069064	SRR12801731
ITM-2018-00101	N3913	Ethiopia	L7	CP069063	SRR12801730
ITM-2018-00102	N0069	China	L1	CP069062	SRR12801729
ASM1292376V1	-	Rwanda	L8	PRJNA598991	SRR10828834
H37Rv	H37Rv	-	L4	AL123456.3	
M. canettii		-	-	HE572590.1	

5.3.3 Genome assembly and annotation

Barcoding and basecalling of long reads was performed with Guppy Basecalling Software v3.2.43 using the command*: flowcell - FLO-MIN106, kit - SQK-LSK109, barcode_kits - "EXP-NBD104 EXP-NBD114"*. De novo assembly was performed with Flye v2.5 [21] under default settings. Assemblies were further improved in a multi-step process: three rounds of polishing with Racon [22] (using long reads) and three rounds of polishing with Pilon [23] (using short reads). Completeness and quality of the assemblies were assessed based on the number of complete BUSCOs (Benchmarking Universal Single-Copy Orthologs) calculated with BUSCO.v3 [24] using the *Actinobacteria* set; and number of contigs. These assemblies were further oriented with Revseq v6.6.0.05 based on the H37Rv reference genome and rotated to equivalent start locations with Circlator fixstart [25]. All assemblies were aligned to the H37Rv reference using nucmer from MUMmer v3.2[26].

The 18 new assemblies as well as three additional assemblies (H37Rv, *M. canettii* and a lineage 8 strain) were annotated using the RAST server [27] under default settings. To validate the accuracy of the annotations of PE/PPE genes we used BLAST to query all the assemblies against a custom H37Rv PE/PPE gene database. Gene annotation was manually curated to ensure the presence of a proper start/stop codon, and proper distance to RBS motif. Genes that failed validation or that were absent from the BLAST search were marked as either pseudogenes or missing. A subset of PE/PPE genes identified by RAST were not given specific gene identities and were only annotated based on the gene family, e.g. 'PE family protein', 'PE-PGRS family protein', or 'PPE family protein'. These unlabeled genes were validated as putative yet unannotated PE/PPE genes via a BLAST search against H37Rv PE/PPE genes to ensure they had not already been annotated previously. They underwent further validation as PE/PPE genes based on passing criteria developed by Ates et al [28]. Namely, we verified the presence of a PE or PD motif within the N-terminus and the presence of YxxxD/E motif for the putative PE genes, and the presence of a PPE motif within the N-terminus and a WxG motif within the N-terminus for the putative PPE genes. Since these motifs are considered hallmarks for PE/PPE genes, we labeled genes that passed both checks as putative new PE/PPE genes.

5.3.4 Sequence analysis

A whole genome phylogeny was made for the 21 *M. tuberculosis* assembly from a concatenated alignment of SNPs provided by an alignment of each assembly to the H37Rv reference using MUMmer's nucmer command. Only sites with variants present in at least eight samples were used to create the concatenated alignment. The tree was generated using FastTree v2.1.10 [29] and rooted using *M. canettii* as an outgroup. We additionally created phylogenies for each of the 168 PE/PPE genes using FastTree on protein alignments performed with MUSCLE v3.8.31 [30].

SNPs output from nucmer were used to determine variations within gene regions from each genome assembly. For PE/PPE genes N-terminal and C-terminal sites were approximated based on findings within a previous study [31]: 0-110 and 0-130 amino acids for PE and PPE proteins respectively. Synonymous and non-synonymous sites were calculated for each PE/PPE protein alignment using SNAP [32] against the H37Rv sequence. Variation across the PE/PPE proteins and lineages was conducted by calculating the pairwise genetic distance for each protein.

Distances were then averaged per lineage and normalized by the total protein sequence length. P values were calculated using fishers exact test.

GenMap (v 1.3.0) [33] was used to determine mappability scores for each genome at default settings. The mappability score ranges from 0 (low mappability) to 1 (high mappability) and is a measure of the repetitiveness of the underlying sequence.

5.4 Results

5.4.1 18 high quality complete M. tuberculosis reference genome assemblies

18 *M. tuberculosis* isolates covering lineages 1-7 were sequenced using nanopore sequencing and then assembled using Flye to generate high quality assemblies (median N50 of 4,424,350 and median genome length of 4,426,106). A single contig was achieved for all but one genome assembly (ITM-2018-00088). Assemblies contained on average 99% the number of genomic features (BUSCOs) expected for the *Actinobacteria* class of which *M. tuberculosis* is a species. In order to juxtapose variations across all lineages of MTB we also included an *M. canetti* strain and a lineage 8 strain. We additionally included a H37Rv strain to validate our results. (Table 5.2). The number of complete genomic features was remarkably lower for the Lineage 8 strain which could possibly be due to its different sequencing procedure (PacBio and Illumina).

Table 5.2 Details on assemblies analyzed in this study

Samples	Lineage	Length of assembly	Number of contigs	GC-content (%)	Genomic features (Quast)*	Complete BUSCOs %**
ITM-2018-00082	L1	4,410,121	1	65.62	4020 + 40	99.43
ITM-2018-00083	L1	4,420,471	1	65.61	4017 + 33	99.43
ITM-2018-00087	L2	4,399,032	1	65.62	4013 + 39	99.15
ITM-2018-00088	L2	4,418,134	2	65.6	4021 + 37	99.43
ITM-2018-00089	L3	4,422,216	1	65.61	4031 + 37	99.43
ITM-2018-00090	L3	4,438,446	1	65.62	4042 + 31	99.43
ITM-2018-00091	L3	4,431,518	1	65.61	4041 + 37	99.15
ITM-2018-00092	L4	4,393,016	1	65.61	4026 + 30	99.43
ITM-2018-00093	L4	4,395,562	1	65.61	4032 + 28	99.15
ITM-2018-00094	L4	4,419,495	1	65.59	4027 + 34	99.15
ITM-2018-00095	L5	4,425,207	1	65.59	4012 + 52	99.15
ITM-2018-00096	L5	4,420,832	1	65.59	4015 + 48	99.43
ITM-2018-00097	L5	4,428,969	1	65.59	4024 + 48	99.43
ITM-2018-00098	L6	4,386,429	1	65.59	3980 + 37	98.86
ITM-2018-00099	L6	4,387,551	1	65.6	3980 + 39	98.86
ITM-2018-00100	L6	4,391,569	1	65.6	3984 + 37	99.43
ITM-2018-00101	L7	4,408,017	1	65.6	4000 + 35	99.43

ITM-2018-00102	L1	4,419,191	1	65.6	4006 + 36	99.43
ASM1292376V1	L8	4,379,910	1	65.62	3988 + 53	92.05
H37Rv	L4	4,411,532	1	65.61	4098 + 0	98.86
M. canetti		4,482,059	1	65.61	3745 + 92	99.43

** complete genomic features + partial*

***Total number of BUSCOs for Actinobacteria odb9 set is 352*

After aligning to the H37Rv genome, we identified 12,254 unique SNPs across the 20 *M. tuberculosis* genome assemblies, 67% (8,264) of which were unique to a single strain. We built a whole-genome phylogeny based on the identified SNPs (Figure 5.1A), which followed the expected clustering of all lineages based on previous studies [5].

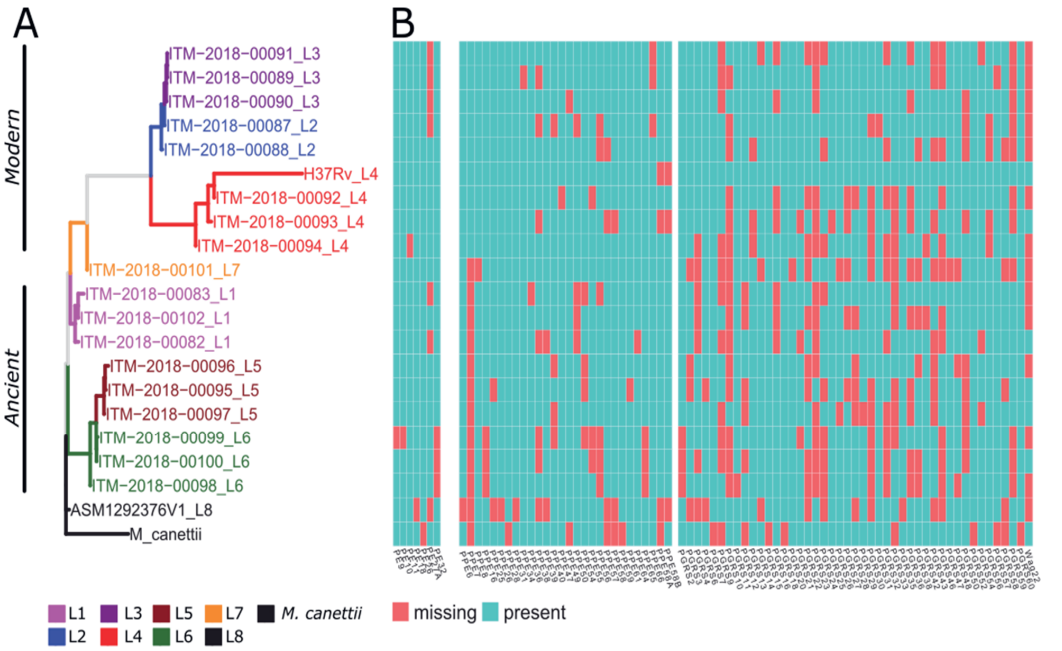


Figure 5.1 Absent / present genes matrix. Panel A shows the whole genome phylogeny of the strains from the eight human-adapted MTBC lineages and *M. canettii*, rooted on *M. canettii* . 'Modern' lineages (2,3,4), and 'ancient' ancestral lineages (1,5,6) form respective clusters in the phylogeny. Lineage 7 is in-between 'modern' and 'ancient' lineages, whereas Lineage 8 is most similar to the ancestral. *M. canettii* is a non-pathogenic relative of *M. tuberculosis*. In panel B the respective presence (blue or absence (pink) is shown for PE (left), PPE (middle) and PE-PGRS (right) genes for each genome assembly in the phylogeny. Only genes with variable presence are included in the figure.

5.4.2 PE/PPE genes show lineage specific gains and losses

We identified 37 PE, 64 PE-PGRS, and 71 PPE genes within our assemblies, including 4 newly annotated genes (PE10A, PE19A, PPE58A, PPE58B). Each gene was present once in 91% of assemblies, however nearly all assemblies contained cases of deleted or non-functional (<50 amino acids)/pseudo PE/PPE genes ((Figure 5.1B). Copies of PPE genes were most frequently

found to be completely absent (deleted) across the genome assemblies, compared to PE and PPE genes. To be exact: 1.8% (14/777), 1.6% (21/1,344) and 2.3% (35/1491) of all PE, PE-PGRS, and PPE gene copies were absent across the genome assemblies respectively. While copies of PE-PGRS genes were most frequently found to be nonfunctional across the genome assemblies. To be exact: 1 (0.12%), 223 (16.6%), and 45 (0.9%) of all PE, PE-PGRS and PPE genes were identified as being nonfunctional, respectively.

We noted that several genes were absent from the majority of genome assemblies. In particular, copies of seven PE-PGRS genes and one PPE gene were absent in half of our genome assemblies (>10). The most frequently absent gene copies were Wag22 (PGRS) and PE-PGRS10; both absent in 14 (67%) and 15 (71%) of genome assemblies respectively. This finding for Wag22 corroborates with previous studies which revealed Wag22 unsuitable for analysis due to deletions and frame shifts in 86% of their samples [7]. The remaining six PGRS/PPE gene copies (PE-PGRS9, PE-PGRS30, PE-PGRS59, PE-PGRS46, PE-PGRS23, PPE7) contained in-frame stop codons that made these proteins non-functional in multiple (10-13) genomes assemblies. It has previously been reported that secretion of PE-PGRS proteins is not essential for the virulence of *M.* tuberculosis [28]. While the large absence of these PGRS gene copies in our data is in line with this assertion, there is little known about the function of these specific genes. The PE27A (28 aa long) was the most frequently absent PE gene in our set (absent in 7 genome assemblies). Due to its limited length, this could have easily been removed as a result of a deletion event.

Presence or absence of certain gene copies may have important biological implications that reflect the efficacy of *M. tuberculosis* to sustain an infection in a host. The clustering of gene presence according to lineage could therefore help to shed light on these biological implications. In our analysis, we observed correlations between gene copy presence and lineage ((Figure 5.1B). We did not observe any gene duplications within these gene families. Overall, Lineage 8 was found to have the most absent gene copies (26), followed by lineage 7 (23), and *M. canettii (*18). The high number of missing gene copies from these lineages is partially a consequence of our data containing only a single genome assembly for these lineages. The remaining lineages were represented by at least two genome assemblies. Amongst these, lineage 6 had by far the highest number of absent PE/PPE gene copies (13). The higher number of absent PE/PPE gene copies from these four lineages can possibly explain differences of virulence when compared to the remaining lineages.

Of all genes, PPE7 was absent from the most lineages (6/9 lineages) and only had functional copies in 3 modern lineages (Lineage 2, 3, 4). The PPE7 protein was found less abundant in hypo-virulent strain H37Ra [34] compared to H37Rv, suggesting this gene might play a role in virulence. It is likely that modern lineages, which are more virulent than ancestral lineages [35], acquired this gene, based on our data showing its absence in ancient lineages. Copies of Wag22, PGRS10 and PGRS59 were nonfunctional in 4 lineages: 2, 3, 7 and 8 (Wag22), 2, 3, 6, 7 (PGRS59) and 1,2,6,7 (PGRS10). Wag22 had previously been reported to be a non-essential gene for in vitro growth of H37Rv [36], which corresponds to its marked absence across several lineages.

5.4.3 Newly discovered PE/PPE genes missing from H37Rv reference annotation

During our analysis, we identified four genes not previously reported as a specific PE/PPE gene, yet possess all the features expected for these genes (Figure 5.2, Table 5.3). We labeled them based on their proximity to neighboring genes from the same gene family: PE10A, PE19A, PPE58A, and PPE58B.

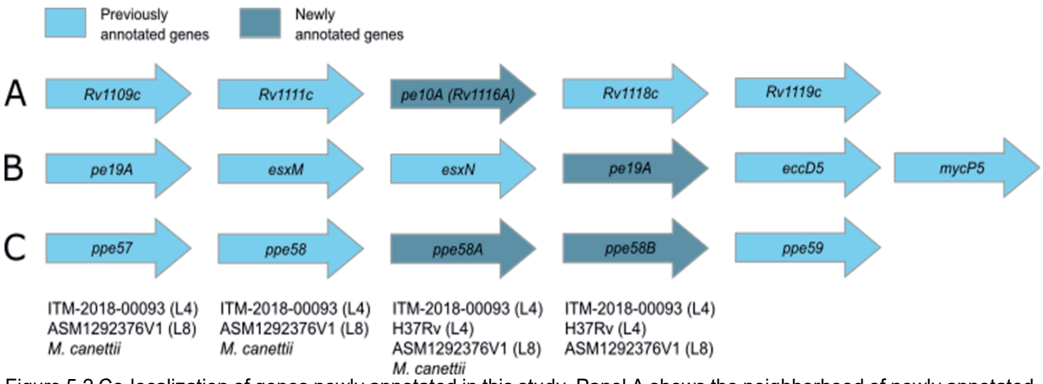


Figure 5.2 Co-localization of genes newly annotated in this study. Panel A shows the neighborhood of newly annotated PE10A gene containing previously annotated genes Rv1109c, Rv1111c, Rv1118c, and Rv1119c. Panel B shows the neighborhood of newly annotated PE19A gene containing the previously annotated pe19A, esxM, esxN, eccD5 and mycP5 genes. Both PE10A and PE19A are present in every genome sequenced in this manuscript. Lastly, panel C shows the neighborhood of newly annotated PPE58A and PPE58B genes, in the context of the previously annotated PPE57, PPE58 and PPE59 genes. In panel C the four left-most genes are annotated with strains from which that particular gene is missing.

Table 5.3: Attributes of PE/PPE genes newly identified in this study

№	Protein name	Annotation according to H37Rv NCBI	Length of the protein	Absent[1]	PE/PPE markers (motif, position)[2]	Substitutions (sample, position)[3]	Amino acid similarity/overlap with other proteins
1	PE10A	Conserved hypothetical protein (fragment)	91 aa	-	PE motif, position 59; YVHAD motif, position 85	M. canettii, position 72 (Y->H)	similarity with C-terminal part of PE17 (68.05% identity in 72 aa alignment, PE17 is 310 aa long)
2	PE19A	Conserved protein	300 aa	-	PE motif, position 31; YVSTD motif, position 38, YGVID motif, position 96	-	No significant (>50% identity to other proteins)
3	PPE58A	None	180 aa	ITM-2018-00093 (L4), ASM1292376V1(L8), H37Rv, *M.canettii*	WIG motif, position 56	ITM-2018-00099 (L6), position 13 (Y->C)	similarity with PPE57 (55% identity in 173 aa alignment, PPE57 is 176 aa long); PPE58 (57% identity in 122 aa alignment, PPE58 is 232 aa long); PPE59 (54% identity in 175 aa alignment,

							PPE59 is 178 aa long)
4	PPE58B	None	131 aa	ITM-2018-00093 (L4), ASM1292376V1(L8), H37Rv	WAG motif, position 25	ITM-2018-00087 (L2), position 84 (E->A); ITM-2018-00089 (L3), position 72 (G->D); M. canettii, multiple substitutions	No significant (>50%) identity to other proteins

[1]Genomes in which this protein is absent

[2]Specific markers used to identify this as a PE or PPE protein

[3]Genomes containing variations of this protein's amino acid sequence

PE10A, annotated as 'PE family protein' by RAST, neighbors PE10 on one side, an inducer of macrophage apoptosis [11] and PE11 on the other side, involved in virulence processes [37]. The three genes appear consecutively in the genome. Although containing features of a typical PE gene (PE motif at position 59 and secretion signal at position 85 YVHAD), it was reported as *Conserved hypothetical protein (fragment)* in the H37Rv reference genome (NCBI accession number AL123456.3). It is well conserved across all the lineages and contains only a single variation in *M. canettii* (Table 5.3, Supplementary Figure 1). Based on this high conservation across lineages and the presence of characteristic PE motifs, we consider PE10A a new PE gene that was previously unannotated within the H37Rv reference genome.

PE19A, which RAST annotated in the H37Rv genome as 'PPE family protein', does not have the expected PPE and WxD motifs inherent to PPE family proteins. Instead, it contains typical PE features (PE motif and YxxxD motifs) (Supplementary Figure 2), leading to its designation as 'PE19A'. Similar to PE10A, it is well conserved across all lineages, including *M. canettii.* Despite being annotated as a PPE family protein, we consider PE19A another previously unannotated PE gene.

Two newly described PPE proteins, PPE58A and PPE58B, are localized together, between PPE58 and PPE59 (Figure 5.2). The two genes both contain a WxG motif (Supplementary Figure 3 and 4). Although PPE58A does not have a PPE motif in the N-terminus, it has an EPP motif in the C-terminus - a feature shared by PPE57, PPE58 and PPE59. Interestingly, the region with PPE58A and PPE58B is missing in H37Rv (L4) despite their presence in other Lineage 4 isolates. In ASM1292376V1(L8) and ITM-2018-00093 (L4) the region containing PPE57, PPE58, PPE58A and PPE58B is absent (Supplement Figure 5b), indicating that this deletion affects a range of lineage 4 strains. *M. canettii* lacks PPE57, PPE58, PPE58A (Supplementary Figure 5d), but interestingly still has PPE58B. Both PPE58A and PPE58B have not been annotated as a gene in the H37Rv reference and are completely absent from the annotation. PPE58B especially shares no significant similarity with any other PE/PPE gene, pointing to its novel nature making it an interesting candidate for further research. The absence of both PPE58A and PPE58B in the H37Rv annotation is a good example of how our high quality assemblies led us to identify genes that would have been otherwise overlooked due to reference bias of the H37Rv genome.

All four of these newly annotated genes exhibit high degrees of conservation across lineages and genome assemblies. Based on their widespread distribution and link to other PE/PPE genes, we

speculate that these genes may play an important role in *M. tuberculosis* virulence and pathogenicity.

5.4.4 PE/PPE Variation Profiles

To explore the increased diversity within PE/PPE genes compared to other genes within *M. tuberculosis*, we analyzed the collection of SNPs identified with nucmer. The majority of the 12,254 unique SNPs were found within coding regions (10,503, 86%). Of these, 993 (8% of total) were located within PE/PPE regions. SNPs were observed at a higher density within PE/PPE genes (average 0.48 snps/KB) compared to other genes (average 0.24 snps/KB). The *M. canettii* genome assembly followed the same trend at an average of 7.67 snps/KB for PE/PPE genes compared to 4.41 for other genes. This correlates with previously reported SNP rates for these regions compared to regular *M. tuberculosis* genes [16].

It has previously been reported that PE/PPE genes undergo a weaker purifying selection compared to other genes, as evaluated through the frequency of synonymous and nonsynonymous mutations [14]. We therefore investigated the rate of nonsynonymous/synonymous substitutions (dn/ds ratios) in PE/PPE genes, where values of <1 denote purifying selection, >1 denote positive selection and =1 denote neutral evolution. We observed greater levels of purifying selection within PE/PPE genes (mean 0.4) when compared to previously reported non PE/PPE dn/ds values (mean 0.57 [16]), indicating that most variation within PE/PPE genes are synonymous. Our values are considerably lower than reported in previous studies examining the selective pressure in PE/PPE genes [14,16]. This discrepancy could be attributed to the increased number of genes surveyed, the higher quality genome assemblies, and the broader range of lineages analyzed. PE genes were under the most purifying selection (mean 0.14) compared to that of PPE (mean 0.46) and PE-PGRS genes (mean 0.4).

To evaluate the effect variability could have on the overall protein structure, we calculated a variability score as previously described [16]. This score represents the proportion of unique amino acid sequences across all 21 genome assemblies – thereby focusing on variations leading to an amino acid change and ultimately impact on protein structure. We compared our findings to the similar analysis by McEvoy et al. [16] (Figure 5.3). Although we included PGRS genes, due to limitations of sequencing, the study of McEvoy et al. was unable to survey PGRS genes, due to the repetitive nature of these genes.

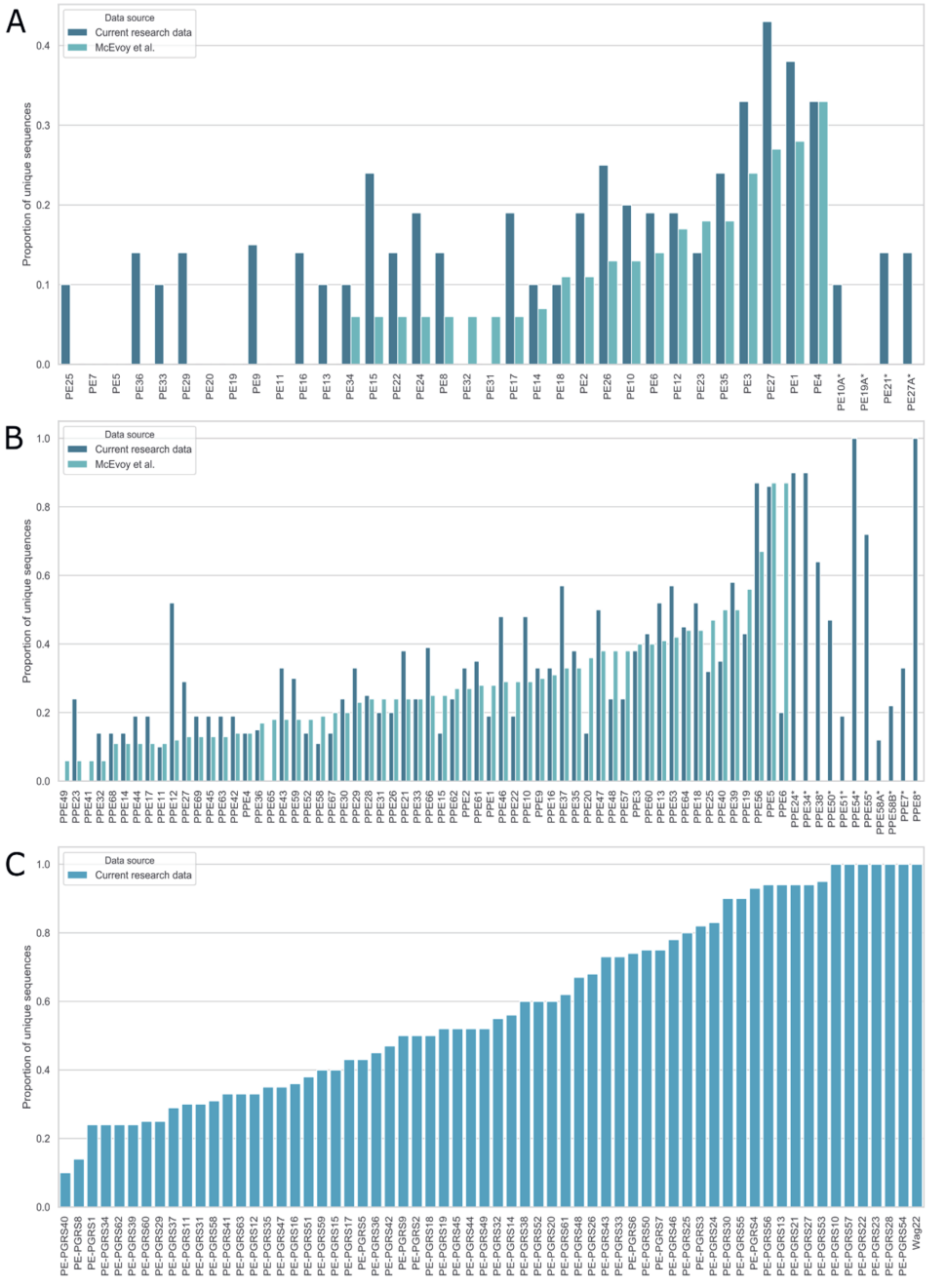


Figure 5.3 Proportion of unique protein sequences for (a) PE, (b) PPE and (c) PE-PGRS proteins. For PE and PPE, levels from this study (blue) were compared with those reported by McEvoy et al (green) [16] The bins are sorted according to variability level reported in McEvoy et al. Asterisks denote a protein not considered in McEvoy et al.. A

proportion of 1 indicates that all amino acid sequences of the protein are unique across the samples containing that gene. Zero denotes all amino acid sequences of that protein are identical across all samples

Overall, our results corroborate and extend those of McEvoy et al. [16]. Discrepancies could be explained by McEvoy's not covering lineages 3, 7, 8 and *M. canettii,* in addition to having only a single representative for lineages 5 and 6.

We observed significantly more conservation in PE genes (median=0.14) than in PE-PGRS (median=0.54, p-value=<0.005) and PPE genes (median=0.34, p-value=<0.005). Complete conservation was observed in 5 genes (PE5, PE7, PE11, PE20, PE19) reported to be essential for localization on mycobacterial cell wall [38]. We also observed PPE genes that exhibited considerable conservation (variability levels equal to 0): PPE65, PPE49, and PPE41. In contrast, six PPE genes (PPE5, PPE24, PPE34, PPE56, PPE8, and PPE54) had variability levels higher than 0.8, indicating they differed in at least 76% (16/21) of the genome assemblies. These genes are relatively large (mean 2230 aa). The highest degrees of variability were observed amongst PE-PGRS genes. Seven PE-PGRS genes (PE-PGRS10, PE-PGRS57, PE-PGRS22, PE-PGRS23, PE-PGRS28, PE-PGRS54, Wag22) exhibited a variability level of 1, indicating each genome assembly had a different sequence of the gene.

Although PE-PGRS genes are generally considered to be among the most variable *M. tuberculosis* genes, making them hard to analyze, we observed two PE-PGRS genes that displayed high conservation: PE-PGRS40 (0.1) and PE-PGRS8 (0.14).

5.4.5 Mappability of PE/PPE genes

PE and PPE genes are typically excluded from sequence analysis studies due to concerns over the alignment calls to these repetitive regions [12]. However, due to the high conservation observed in a subset of these PE/PPE genes, we hypothesized that some genes need not be excluded from analysis. This was confirmed after calculating whole genome mappability scores for each of the assemblies (Supplementary Table 2). Mappability scores are a measure of the repetitiveness of k-mers for each position in the genome and were calculated using GenMap (See Methods). A mappability score of 1 indicates a completely unique k-mer, whereas low mappability scores indicates k-mers in repetitive regions which are more difficult to map. Mean mappability scores within PE genes matched that of non PE/PPE regions of the whole genome (0.98 PE vs an average of 0.98 across the complete genomes). PE-PGRS and PPE genes had the lowest mappability scores at 0.86 and 0.9 respectively. Despite low mappability scores for a subset of the PPE/PE-PGRS genes, 6 PE-PGRS and 25 PPE genes had scores greater than 0.98 (Figure

5.4). Our results suggest there is no need to exclude all PE/PPE genes from *M. tuberculosis* comparative genomics studies.

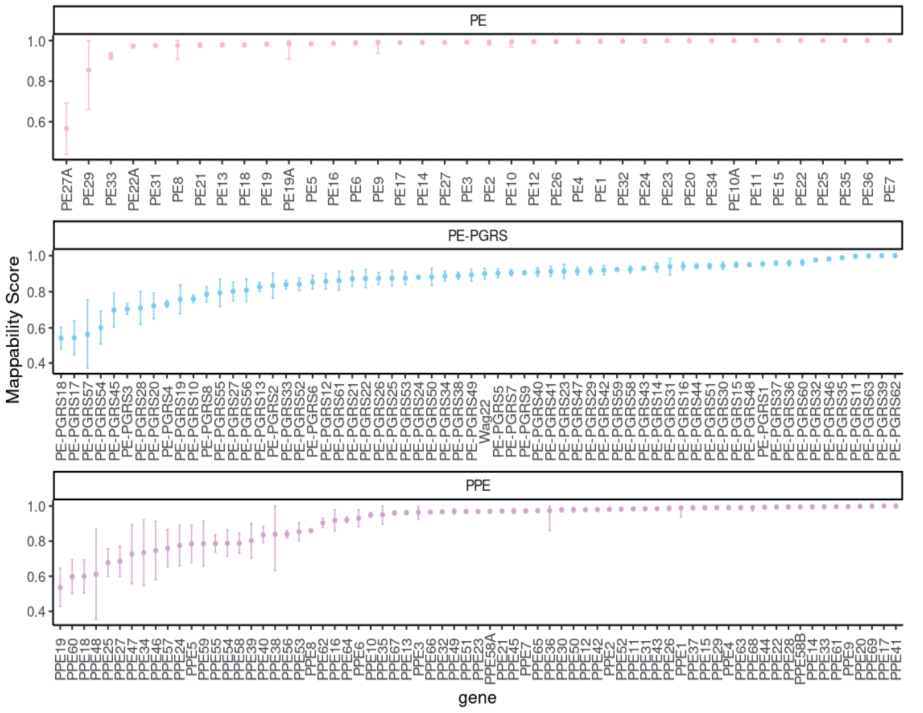


Figure 5.4 Mean mappability scores for PE/PPE genes. Mappability scores were computed for each assembly using the GenMap program and give an indication to how well each gene can be mapped during alignment analysis. Scores are averaged per gene region across all genomes. Standard deviation bars show the range across genome averages.

5.4.6 Lineage based variation profiles

As different *M. tuberculosis* lineages can differ in their ability to induce and transmit disease, it is important to explore genetic variation between and across lineages. To better understand the connection between protein variation and lineage, we calculated the pairwise protein distance for every gene across the 21 genome assemblies and normalized by the length of the multiple protein alignment (see Methods; Figure 5.5, Supplementary Figure 6, Supplementary Figure 7). This allowed us to identify genes that tend to vary greatly within a lineage versus those that contain lineage specific variations. As expected, proteins exhibited higher protein distances on average between genome assemblies of different lineages (mean=0.1, Supplementary Figure 6) than genome assemblies within the same lineage (mean=0.06, Supplementary Figure 7). Interestingly, we also observed considerable variation for genes within the same lineage for 11 genes (> 30% variability, 10 PE-PGRS, 1 PPE genes). PE genes were particularly conserved within genome assemblies of the same lineage (87% of genes with < 5% variability) compared to PE-PGRS (39% of genes) and PPE (71% of genes) genes. Likewise, PE genes observed the most conservation

between lineages (76% of genes < 0.05 variability) compared to PE-PGRS (27% of genes) and PPE (63% of genes) genes.

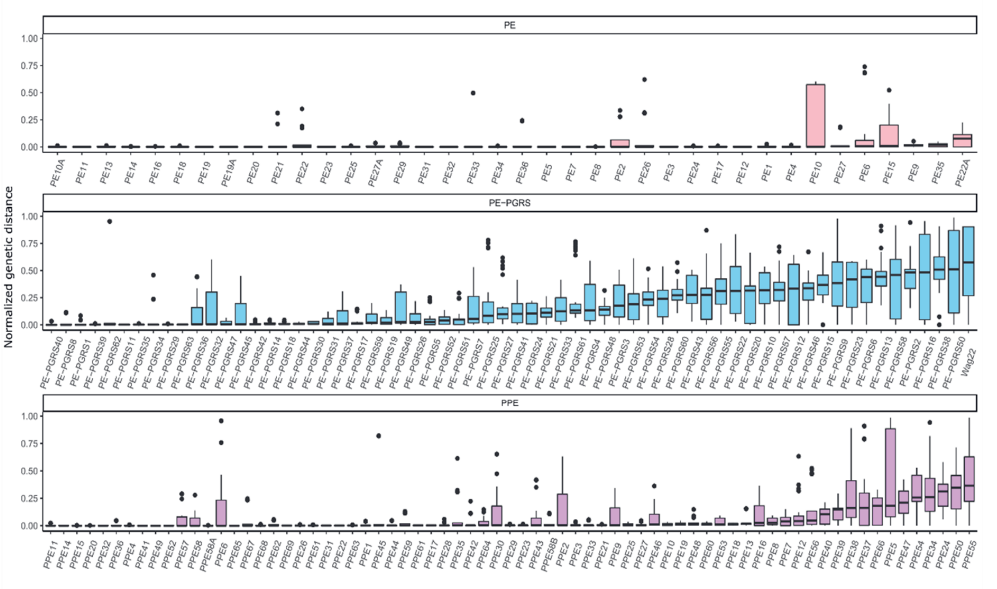


Figure 5.5 Normalized lineage distance per protein. The Y-axis represents the distance between the samples per protein normalized by the length of MSA. Boxplots display the distribution of the distances with the median value marked with a horizontal line. The colors indicate the protein family, i.e., PE (pink), PPE (sky-blue), PE-PGRS (purple).

We observed a number of genes with particularly high intra and inter-lineage variability (Supplementary Table 1). For example, PE10 had the highest intra- and inter-lineage variation across the PE family. PE-PGRS proteins, specifically PE-PGRS2 and PE-PGRS38, were generally highly variable both across and within lineages due to large deletions in its C-terminal domain. Amongst the PPE gene family, PPE24 and PPE47 both had highest levels of inter/intra-lineage variability.

5.4.7 Lineage informative genes

We identified genes whose variability were specific to certain lineages by finding genes that exhibited high pairwise inter-lineage protein variations for one lineage but conservation in pairwise variations across remaining lineages. Such genes could explain characteristics of *M. tuberculosis* that are dependent on lineage. We observed genes exhibiting lineage specific variation across each PE/PPE gene family. PE-PGRS62, PPE45, PE6, and PE10 genes exhibited the highest degree of lineage specificity (as calculated by the difference of pairwise protein distances between the variable lineage and the remaining lineages) (Figure 5.5). Interestingly, these genes were all specific for different lineages.

PE-PGRS62 exhibited variation only within lineage 7 (Figure 5.6). It is described to be a T cell antigen [39] and was reported to be extremely conserved, leading to its previous investigation as a possible TB vaccine candidate [16,40]. However, in our data, the lineage 7 isolate has a very short protein (64 aa), which might be the result of a frameshift and should be considered during the possible development of a vaccine. Our results were able to elucidate this lineage specific variation for this gene, as previous studies considering PE-PGRS62 as a potential vaccine candidate did not include lineage 7 in their analysis.

PE10 was nearly conserved in all but two lineages (L2 and L3), which suggests this gene contributes to the division of this clade to other lineages of *M. tuberculosis*. Lineage 2 and Lineage 3 representatives of PE10 were different in length compared to other lineages and other samples from the same lineage (Figure 5.6). These lineages are localized at the same clade and belong to the 'modern', more virulent, lineages. PE9-PE10 complex was described to participate in macrophage apoptosis [41]. Such variability of PE10 in Lineage 2 and 3 could be the result of the possible adaptation for better infection of the host organism.

PE6 was completely conserved in all human adapted lineages of *M. tuberculosis* (Figure 5.6), only showing variation compared to the *M. canettii* strain. Because very little is known about the function of PE6 or its role in MTB pathogenicity, this gene could be an interesting target of future research studies. Lastly, PPE45 exhibited variation only specific to lineage 6. Similar to PE6 there is not much known about this gene.

Amongst all *M. tuberculosis* lineages, lineage 2 has been widely reported to be one of the most virulent, exhibiting a mild inflammatory response[42–44]. It could be the case that members within the PE/PPE gene family contribute to the virulence of this lineage. Amongst all the genes with lineage specific variation, PE-PGRS4 exhibited the highest variation specific to solely lineage 2. As little is known about this gene further studies into its role in virulence could shed light on the mild inflammatory response noted within lineage 2 patients.

Figure 5.6 Normalized pairwise protein distance for each lineage computed for select genes (PE-PGRS62, PE10, PE6 and PPE45). Distance values were calculated by averaging pairwise protein distances across each representative genome for a lineage. Distances were then normalized based on the length of the protein (multiple sequence alignment). Lineages with only a single representative genome were given NA as a distance value, here shown as grey boxes.

5.5 Discussion

Here we present 18 reference genome assemblies covering all major *M. tuberculosis* lineages based on nanopore long read and Illumina short read sequencing data. These complete and extremely high-quality genome assemblies enable future studies on MTB genomics in many ways. We contribute several complete genome assemblies for lineages 5, 6 and 7 which are a rare resource. These assemblies will make it easier to characterize and understand the nuances between the different lineages of *M. tuberculosis*. Having high quality assemblies from different lineages also aids to combat the general reference bias towards the H37rv strain with lineage specific references. Lastly, thanks to the span of long read data, each of the assemblies has high contiguity, even for difficult to assemble regions of *M. tuberculosis* such as the highly repetitive PE/PPE genes.

PE/PPE genes have a number of special properties that make them an interesting point of study. They are unique to mycobacteria and account for 10% of the coding potential of *M. tuberculosis*, suggesting that they play an important role in the overall function of *M. tuberculosis*. These genes are surface associated cell wall proteins which evidence suggests could provide a diverse antigenic profile and affect immunity[45]. Additionally they have been reported to be good candidates for MTB vaccines, play a role in bacterial pathogenicity, and influence host immune response [46,47].

However, despite their apparent importance, PE/PPE genes are typically excluded during TB studies due to the difficulty in mapping regions within their repetitive C-terminal domains [12] , making their function not fully understood. By utilizing both our *de novo* assembled genome assemblies as well as three additional publicly available assemblies (H37Rv, *M. canetti*, and a lineage 8 strain), we were able to resolve and comparatively analyze all 172 PE/PPE genes within and across all lineages. To our knowledge, this is the first study to analyze all PE/PPE genes across all TB lineages. PE genes were found to be the most conserved family with few variations across lineages. This finding reflects that of previous work [16] and can be explained by the absence of the largely variable C-terminal domains. The PE-PGRS subfamily was the most variable with the least amount of known functional properties [7].

We identified previously unreported members of PE/PPE families, namely, PE10A, PE19A, PPE58A, PPE58B. These proteins are highly conserved across all the lineages and *M. canettii*, pointing to a possible important role in MTB. PE10A and PE19A had been identified in the H37Rv reference annotation but not as a PE/PPE protein. More interesting, PPE58A and PPE58B were completely absent from the H37Rv genome, explaining why these genes have not been investigated in previous studies about PE/PPE genes. Further studies are needed to provide a more comprehensive characterization of their function and utility to *M. tuberculosis.*

On the whole, we observed trends between proteins identified as being highly variable within our study and that of previous research. We report greater levels of purifying selection within PE/PPE genes than compared to non-PE/PPE genes. Indicating that a large portion of gene variation within these genes are synonymous mutations. There was also evidence of PE/PPE genes that exhibited lineage specific variation. These genes could explain lineage specific differences of pathogenicity and virulence observed within *M. tuberculosis*. Additionally, we characterized PE/PPE genes based on their mappability to a reference genome and found the majority of PE genes and many PE-PGRS/PPE genes could confidently be aligned to a reference genome, thereby removing the commonly held practice of excluding all PE/PPE genes from analysis.

Previous studies have pointed to the possible use of PE/PPE genes as potential vaccination targets. [48] These genes have been identified to have several characteristics which could aid in vaccination studies: localization at cell surface, interaction with mitochondria driven necrosis of cells, and their binding potential to immunologically active receptors such as the TLR2 receptor[48]. By analyzing variations present across and within different lineages, we identified potential vaccine targets based on gene conservation across these lineages. The function of PPE51 protein is unknown[13], however, it was extremely conserved according to previous studies [16], and our study confirmed this result. Therefore, it can be considered as a good candidate for further vaccine investigations. PPE69 was also very conserved, with only one amino acid variation in one sample of Lineage 6 and *M. canettii*. The function of this protein has not been reported anywhere, to the best of our knowledge. Among other highly conserved candidates with unknown function were PE7, PE18, PE32, PE33, PE34, PE36, PE-PGRS1, PE-PGRS8, PE-PGRS40.

In addition to new potential vaccine targets, we identified previously identified potential vaccine targets, that should be reconsidered based on higher variation across lineages. For example, Xu et al considered PPE57 as a candidate based on its ability to induce Th1-type immune response

activity in mice [49]. However, our results show variation among Lineage 1 samples and the entire absence of this protein in Lineage 8 and *M. canettii*. Due to its variability and absence in some isolates, it might not be suitable for a unified vaccine. Another example is PPE18, which is an important subunit of the M72/AS01E vaccine candidate [50]. According to the WHO [1], this candidate was successful in a Phase IIb trial among patients with evidence of latent TB infection, and if a Phase III trial shows the same results, this vaccine might be a new effective tool to prevent TB. However, our results showed some variability of PPE18 among all lineages, which was also reported by Homolka et al. [51]. Therefore, it is important to investigate the effectiveness of this vaccine carefully across all the lineages.

Overall, we identified several genes, e.g., PPE51, PPE69, which can be putative candidates for vaccination studies based on their high degree of conservation within our study. We also highlight certain PE/PPE genes that are less suitable as a vaccine target due to their variability or absence in specific lineages (PPE57, PPE18). This study highlights PE/PPE genes that should be either considered or excluded as potential vaccine studies based on their variability across our diverse data set.

To conclude, understanding the dynamics and functionality of members within the PE/PPE gene family is still an active and ongoing process. Because these genes make up nearly 10% of the coding potential of the *M.* tuberculosis genome, understanding these genes and their role in pathogenicity and immune response is highly relevant. The data presented in this study forms a baseline in the characterization of these important genes from a genomics perspective. The results and insights found in this study will be instrumental to advance future studies of these regions. Further biochemical studies to expand on the function of these genes could open the door to better understanding of TB virulence pathogenicity and immune response and assist in future vaccine development efforts to combat this deadly bacterial pathogen.

5.6 References

1. Global Report 2019.

2. Achtman M. Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens. Annu. Rev. Microbiol. 2008;

3. Portevin D, Gagneux S, Comas I, et al. Human Macrophage Responses to Clinical Isolates from the Mycobacterium tuberculosis Complex Discriminate between Ancient and Modern Lineages. PLoS Pathog. 2011; 7:e1001307

4. Coll F, McNerney R, Guerra-Assunção JA, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat. Commun. 2014; 5:

5. Ngabonziza JCS, Loiseau C, Marceau M, et al. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. Nat. Commun. 2020;

6. Maiden MCJ, Bygraves JA, Feil E, et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. 1998; 95:3140–3145

7. Copin R, Coscollá M, Seiffert SN, et al. Sequence diversity in the pe_pgrs genes of Mycobacterium tuberculosis is independent of human T cell recognition. MBio 2014; 5:

8. Delogu G, Brennan MJ, Manganelli R. PE and PPE Genes: A tale of conservation and diversity. Adv. Exp. Med. Biol. 2017; 1019:191–207

9. Bottai D, Di Luca M, Majlessi L, et al. Disruption of the ESX-5 system of Mycobacterium tuberculosis causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. Mol. Microbiol. 2012; 83:1195–1209

10. Daim S, Kawamura I, Tsuchiya K, et al. Expression of the Mycobacterium tuberculosis PPE37 protein in Mycobacterium smegmatis induces low tumour necrosis factor alpha and interleukin 6 production in murine macrophages. J. Med. Microbiol. 2011; 60:582–591

11. Tiwari B, Ramakrishnan UM, Raghunand TR. The Mycobacterium tuberculosis protein pair PE9 (Rv1088)-PE10 (Rv1089) forms heterodimers and induces macrophage apoptosis through Toll-like receptor 4. Cell. Microbiol. 2015; 17:1653–1669

12. Galagan JE. Genomic insights into tuberculosis. Nat. Rev. Genet. 2014; 15:307–320

13. . Mycobacterium Tuberculosis: Molecular Infection Biology, Pathogenesis, Diagnostics and New Interventions. Mycobacterium Tuberc. Mol. Infect. Biol. Pathog. Diagnostics New Interv. 2019;

14. Phelan JE, Coll F, Bergval I, et al. Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. BMC Genomics 2016; 17:1–12

15. Bainomugisa A, Duarte T, Lavu E, et al. A complete high-quality MinION nanopore assembly of an extensively drug-resistant Mycobacterium tuberculosis Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. Microb. Genomics 2018; 4:

16. McEvoy CRE, Cloete R, Müller B, et al. Comparative analysis of mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. PLoS One 2012; 7:

17. Gutacker MM, Mathema B, Soini H, et al. Single-Nucleotide Polymorphism–Based Population Genetic Analysis of Mycobacterium tuberculosis Strains from 4 Geographic Sites . J. Infect. Dis. 2006; 193:121–128

18. Borrell S, Trauner A, Brites D, et al. Reference set of Mycobacterium tuberculosis clinical strains: A tool for research and product development. PLoS One 2019; 14:e0214088

19. Camus JC, Pryor MJ, Médigue C, et al. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiology 2002;

20. Supply P, Marceau M, Mangenot S, et al. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. Nat. Genet. 2013;

21. Kolmogorov M, Yuan J, Lin Y, et al. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 2019; 37:540–546

22. Vaser R, Sović I, Nagarajan N, et al. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;

23. Walker BJ, Abeel T, Shea T, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One 2014; 9:e112963

24. Simão FA, Waterhouse RM, Ioannidis P, et al. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;

25. Hunt M, Silva N De, Otto TD, et al. Circlator: Automated circularization of genome assemblies using long sequencing reads. Genome Biol. 2015;

26. Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes. 2004; 5:

27. Aziz RK, Bartels D, Best A, et al. The RAST Server: Rapid annotations using subsystems technology. BMC Genomics 2008;

28. Ates LS. New insights into the mycobacterial PE and PPE proteins provide a framework for future research. Mol. Microbiol. 2020;

29. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 2009; 26:1641–50

30. Edgar RC. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004;

31. Gey Van Pittius NC, Sampson SL, Lee H, et al. Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. BMC Evol. Biol. 2006;

32. . Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. Mol. Biol. Evol. 1994;

33. Pockrandt C, Alzamel M, Iliopoulos CS, et al. GenMap: Ultra-fast computation of genome mappability. Bioinformatics 2020;

34. Zheng H, Lu L, Wang B, et al. Genetic basis of virulence attenuation revealed by comparative genomic analysis of Mycobacterium tuberculosis strain H37Ra versus H37Rv. PLoS One 2008;

35. Krishnan N, Malaga W, Constant P, et al. Mycobacterium tuberculosis lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. PLoS One 2011;

36. Espitia C, Laclette JP, Mondragón-Palomino M, et al. The PE-PGRS glycine-rich proteins of Mycobacterium tuberculosis: A new family of fibronectin-binding proteins? Microbiology 1999;

37. Singh P, Rao RN, Reddy JRC, et al. PE11, a PE/PPE family protein of Mycobacterium tuberculosis is involved in cell wall remodeling and virulence. Sci. Rep. 2016;

38. Cascioferro A, Delogu G, Colone M, et al. PE is a functional domain responsible for protein translocation and localization on mycobacterial cell wall. Mol. Microbiol. 2007;

39. Chaitra MG, Shaila MS, Nayak R. Characterization of T-cell immunogenicity of two PE/PPE proteins of Mycobacterium tuberculosis. J. Med. Microbiol. 2008;

40. Namouchi A, Karboul A, Fabre M, et al. Evolution of Smooth Tubercle Bacilli PE and PE_PGRS Genes: Evidence for a Prominent Role of Recombination and Imprint of Positive Selection. PLoS One 2013;

41. Tiwari B, Ramakrishnan UM, Raghunand TR. The Mycobacterium tuberculosis protein pair PE9 (Rv1088)-PE10 (Rv1089) forms heterodimers and induces macrophage apoptosis through Toll-like receptor 4. Cell. Microbiol. 2015;

42. Reiling N, Homolka S, Walter K, et al. Clade-specific virulence patterns of Mycobacterium tuberculosis complex strains in human primary macrophages and aerogenically infected mice.

MBio 2013;

43. Via LE, Weiner DM, Schimel D, et al. Differential virulence and disease progression following mycobacterium tuberculosis complex infection of the common marmoset (callithrix jacchus). Infect. Immun. 2013;

44. Tram TTB, Nhung HN, Vijay S, et al. Virulence of Mycobacterium tuberculosis Clinical Isolates Is Associated With Sputum Pre-treatment Bacterial Load, Lineage, Survival in Macrophages, and Cytokine Response. Front. Cell. Infect. Microbiol. 2018;

45. Voskuil MI, Schnappinger D, Rutherford R, et al. Regulation of the Mycobacterium tuberculosis PE/PPE genes. Tuberculosis 2004;

46. Fishbein S, van Wyk N, Warren RM, et al. Phylogeny to function: PE/PPE protein evolution and impact on Mycobacterium tuberculosis pathogenicity. Mol. Microbiol. 2015;

47. Brennan MJ. The Enigmatic PE / PPE Multigene Family Vaccination. Infect. Immun. 2017;

48. Brennan MJ. The PE/PPE Multigene Family of Mycobacteria and TB Vaccines. Adv. Tech. Biol. Med. 2017;

49. Xu Y, Yang E, Huang Q, et al. PPE57 induces activation of macrophages and drives Th1-type immune responses through TLR2. J. Mol. Med. 2015;

50. Van Der Meeren O, Hatherill M, Nduba V, et al. Phase 2b Controlled Trial of M72/AS01 E Vaccine to Prevent Tuberculosis . N. Engl. J. Med. 2018;

51. Homolka S, Ubben T, Niemann S. High sequence variability of the PPE18 gene of clinical Mycobacterium tuberculosis complex strains potentially impacts effectivity of vaccine candidate M72/AS01E. PLoS One 2016;

6 Computational methods for strain-level bacterial detection in DNA sequencing data

6.1 Abstract

Metagenomic sequencing is a powerful tool for examining the diversity and complexity of microbial communities. Most widely-used tools for taxonomic profiling of metagenomic sequence data allow for a species-level overview of the composition. However, individual strains within a species can differ greatly in key genotypic and phenotypic characteristics, such as drug resistance and virulence. Therefore, the ability to resolve microbial communities down to the level of individual strains within a species is critical to interpreting metagenomic data for applications in epidemiology and the clinic, where identifying a particular strain, or tracking a particular strain across a set of samples, can help aid in diagnosis, treatment, or the identification and containment of an outbreak.

Recently published approaches have begun to tackle the problem of resolving strains within a particular species in metagenomic samples. In this review, we present an overview of these new algorithms and their uses, including methods based on assembly reconstruction, reference alignments, or patterns. While existing metagenomic analysis methods show reasonable performance at the species and higher taxonomic levels, they perform sub-optimally when identifying closely-related strains within a species. There is a clear need for more fine-grained, high-resolution algorithms to analyze strain-level diversity.

6.2 Introduction

Within a species, bacteria can be highly diverse in terms of their virulence, resistance to antibiotics, geographical transmission patterns, and other phenotypic characteristics[1,2]. Individual strains can vary greatly with respect to pathogenicity, treatment options, transmissibility, and growth rate [3,4]. In order to effectively treat patients, study bacterial population dynamics, conduct epidemiological surveillance, and stem outbreaks, it is critical to identify which specific strains of a species present in a sample [1,5]. Tracking and comparing individual strains shared across sets of samples would allow for the assessment of the evolution of population diversity in longitudinal samples within a patient or other host system. The ability to identify specific strains in a noisy background of other organisms present in a metagenomic sample could allow for improved tracking of strains involved in an outbreak across a population.

Importantly, accurately identifying specific pathogenic strains would aid in patient diagnosis, allowing for personalized treatment regimens, improved treatment outcomes, and a reduction in the spread of antibiotic resistance. To add further complexity, the ability to identify the presence of multiple specific strains of a pathogen would further help with patient treatment and outcomes. Mixed infections, defined as infections caused by multiple strains of a single pathogen species [6,7], represent an underappreciated challenge to understanding infections and have been described for at least 22 bacterial species [3], including *M. tuberculosis* [6,8], *C. difficile* [9,10],

and *Streptococcus pneumoniae* [11,12]. It is estimated that 10-20% of *M. tuberculosis* patients in high risk areas [8,13,14] and 10% of *Staphylococcus aureus*[15,16] patients are infected with multiple pathogenic strains. Mixed infections put patients at a higher risk of treatment failure [3,6,8], as strains with different drug susceptibility and antibiotic resistance profiles [17,18] can complicate diagnosis and identification of the optimal treatment regimen [3]. In addition to poor treatment outcomes, mixed strain infections can increase pathogen virulence due to selective pressure within the host [19]. Accurate classification of individual strains is critical for identifying mixed infections and will help determine proper treatment options for patients with complex infections, track transmission of pathogenic strains in a population, and differentiate between reinfection and intra-host pathogen evolution.

The challenges of identifying specific strains of a species can be divided into two main categories: (1) Detecting individual strains when the constituents of a sample are all members of the same species (i.e. in a cultured microbial sample, or for tracking within-host evolution of a pathogen); (2) Detecting individual strains of a species in a sample containing a mixture of species, such as a metagenomic sample. This literature review aims to cover these applications in discussing the different computational and statistical methods for strain-level inference from cultured and batch WGS data. While there is clearly substantial value in being able to pinpoint individual strains within metagenomic samples, most current widely-used tools for metagenomic analysis only allow for an assessment of composition at the genus or species level, not the strain level. For example, the current most popular metagenomics taxonomic classification programs, including Kraken [20] and MetaPhlAn2 [21], are capable of identifying mixed populations only at the species or genus level--not at the individual strain level within a species. Tools capable of conducting classification of metagenomic samples for higher taxonomic levels such as the family, genus, or species have been previously reviewed [22–25].

In contrast, tools to detect taxonomy at a finer-grained taxonomic levels within metagenomic samples -- targeting specific strains within a species -- are still in their infancy [26]. Most algorithms for strain-level detection have been published within the past five years. To date, there have been no reviews focused on strategies to computationally classify heterogeneous bacterial populations using WGS data at the level of specific strains within a species. This literature review gives an overview of recent methods for classification at the intra-species, or strain level, including methods based on WGS data to identify both specific strains, as well as mixes of strains. These tools are divided into assembly based, alignment-based, pattern based, and reference free methods. We have included both secondary sources (reviews or methods papers) and original research, where the main objective is developing a novel methodology for detecting heterogeneous bacterial communities, e.g. mixed infections or within host evolution. Although we focus on clinical applications here, the methods discussed are applicable to a broad range of biological scenarios, including evaluation of soil, wastewater or other bacterial communities. We discuss appropriate applications of each strategy, evaluation of these strategies in literature, as well as the applicability of these algorithms to health and disease.

6.3 Approaches for detecting individual strains of bacteria within a species

Currently available approaches to classifying genetically distinct populations from a sequencing read set can be binned into four categories (Table 6.1) employing: (1) assembly of distinct haplotypes within the sample (assembly based), (2) alignment of reads to reference genome(s) (alignment based), (3) analysis of different patterns present in the data (pattern based) and (4) reference free approaches that rely on applying statistics directly to allele frequencies.

6.3.1 Assembly based approaches

Assembly based approaches attempt to identify individual strains in a mixture by performing genomic assembly, drawing on tools developed for haplotype reconstruction in diploid samples. Figure 6.1 gives an overview of how a read set can be resolved into a set of distinct individual strains using an assembly based procedure.

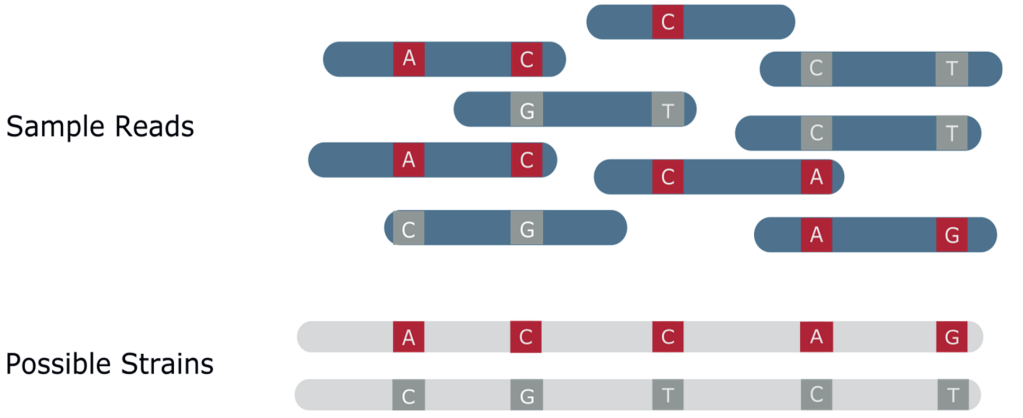


Figure 6.1 **Assembly of multiple distinct strains from a read set**. The blue areas in the sample reads represent regions where the strains have identical sequence. Variant locations in the reads are denoted as red or dark grey stripes. Red variants originate from one haplotype, whereas dark gray variants originate from the other. The goal of an assembly based method is to resolve distinct strains based on the coverage and distribution of the read data, drawing on methods previously developed for resolving haplotypes.

To obtain an accurate reconstruction there must be a sufficient number of sites that differ between the component strains in order to separate or cluster variants into distinct strains [27][28]. Therefore, accurate reconstruction of distinct strains requires sufficient read length to capture overlap between reads, enough discriminating sites to separate populations, and the presence of at least one variant site in most reads.

EVORhA, one of the few assembly based methods designed for reconstructing (assembling) complete bacterial genomes from bulk metagenomic sequencing data, identifies strains via local haplotype assembly (Table 6.1) [29]. For each genomic region containing a sufficient amount of genetic variation, candidate strains are first defined as individual genetically distinct combinations of polymorphisms. To filter out candidate strains that are actually sequencing errors, a minimum number of reads must support an initial candidate strain. In an extension step, candidates are merged with nearby locally constructed candidate strains, based on read frequency and overlap of polymorphism combinations. Ultimately, a mixture model is used to group extended candidate strains occurring at similar frequencies and match these together on a genome-wide level, making the read frequency ratios of observed candidate strains crucial to this method. However, this read frequency criteria for merging strains can produce chimeric strains due to the presence of subpopulations with similar frequencies, similar to a key problem encountered in phasing with whole genome assembly. Given very high coverage, sufficient frequency diversity and sufficient segregating sites, assembly based methods such as EVORhA can resolve the full genomes of genetically distinct subpopulations and yield the most accurate strain identification results when compared to other categories of strain-level identification tools.

Recently, a number of studies have applied long read sequencing data from third generation sequencing platforms such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) to assemble individual strains within metagenomic communities[30,31]. Longer reads could theoretically span the entirety of a strain, making it an attractive option for metagenomic assembly. Somerville et al [32] used the long-read assembler, Flye [33], to reconstruct individual contigs from a long read metagenomic sample, followed by a phylogenetic analysis using NCBI RefSeq to determine strain identity. This was only shown possible for a specific metagenomic community with low complexity, in addition to some manual processing of contigs by the authors. The use of longer reads can mitigate some of the limitations of short-read assembly, allowing for the resolution of difficult to sequence regions and longer contigs. However this comes at the expense of increased errors, lower coverage and higher cost.

Knowing the full sequences of organisms within a sample then allows for comparison and tracking of strains at the highest resolution possible. As such, these methods would be suitable for observing a strain's evolutionary trajectory as well as detecting mixed infections composed of strains that are highly similar to each other. In order to estimate frequencies, a method would need to account for relative abundance of reads specific to each strain. DESMAN [34] does this by exploiting differences in read coverage between genes conserved within a species and other parts of the genome.

A major drawback of assembly based methods is that a large amount of coverage, 50-100× for each strain, is required to achieve an accurate reconstruction, demanding extremely high depth sequencing for strains at a low abundance within a sample [35]. High levels of coverage are required to account for errors introduced by sequencing: each distinct strain must be sequenced with sufficient coverage in order to differentiate spurious variation from true distinct strains. Such high coverages can be achieved in studies where sample complexity is low, with typically less than 5 strains present. Not surprising, this is usually the targeted use case scenario for assembly based metagenomic methods.

6.3.2 Alignment based approaches

Alignment based methods classify strains by aligning reads to a predefined set of reference genomes and applying probabilistic models to calculate a statistical measure representing the likelihood a specific read is associated with a given reference (Figure 6.2) (Table 6.1) [36]. These methods are often considerably faster than assembly based methods and require less coverage, some methods as low as <1×. These methods can achieve such low coverages compared to assembly based methods due to their use of a reference database – where the most likely candidate is selected based on the available data using the probabilistic model. Because alignment based methods share the same limitations, we will discuss the limitations of these methods on the whole at the end of this section.

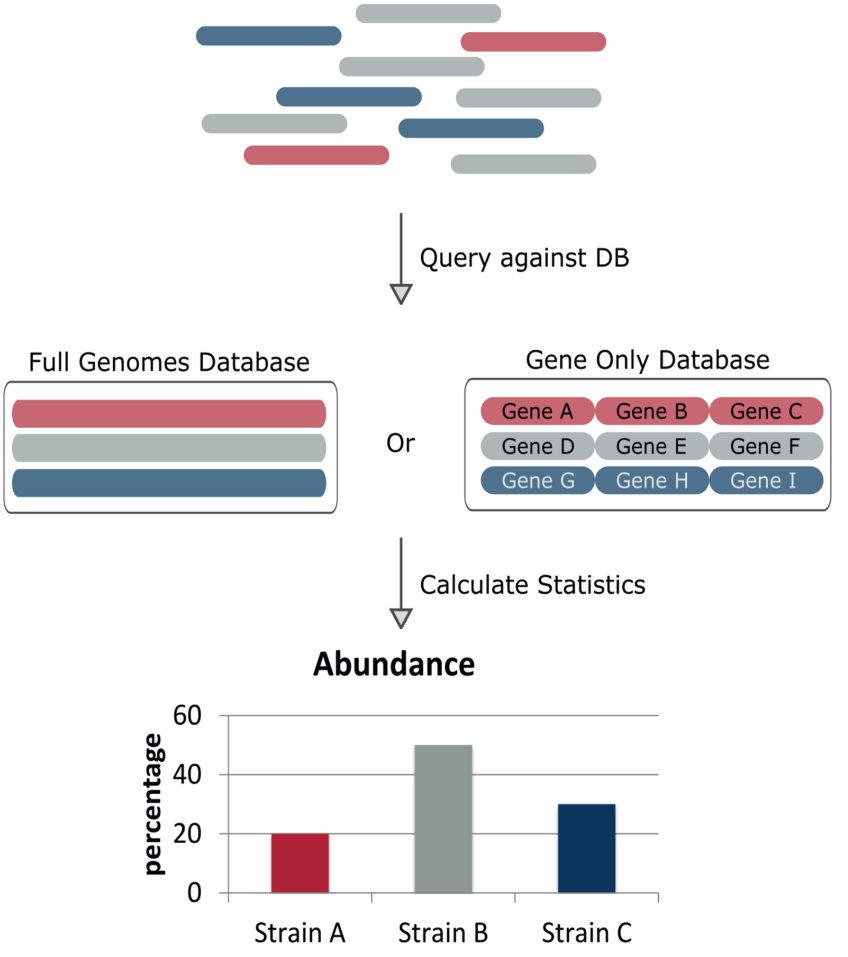


Figure 6.2 **Alignment based approaches** Reads of a sequencing dataset – where different colors denote genetically distinct strains – are aligned to a reference database of full genomes or taxonomic markers (in

this case genes). Strain abundances can be estimated by the relative number of reads aligning to each reference genome.

Pathoscope, [37] one of the most commonly used classification pipelines for metagenomic analysis, uses different aligners (GNUMAP[38], Bowtie 2 [39] and BLAST[40]) to align reads to reference genomes. Scores for each alignment are converted to posterior probabilities that represent the likelihood that an alignment is the source of the read. Non-unique reads are reassigned to their nearest reference using a Bayesian mixture model which uses both the mapping scores and the proportions of non-unique reads.

Another alignment based method, Sigma, allows users to choose their own short-read alignment algorithm, using Bowtie2 as a default [41]. Instead of using scores given by an aligner, Sigma computes its own probability scores for each read to originate from an alignment by examining the number of matches and mismatches between the two.

While most metagenomic methods focus on short-read data, MetaMaps [42] is specifically developed to capitalize on the advantages of long-read sequencing data. MetaMaps uses its own mapping algorithm to align long reads to genomes in a database.

Calculation of strain abundance in alignment-based approaches leverages the number of reads mapping to each reference genome. For Sigma the relative abundance of a genome is simply the proportion of aligned reads out of the total number of reads, whereas Pathoscope calculates relative abundance from the sum of the probability of reads mapped to different genomes in the reference database. The alignment-based strain-identification algorithm BIB exploits the similarities between alignment-based strain identification and the more well-established field of RNA-seq data analysis [39,43–45] for calculating relative abundances, by implementing the RNA-seq algorithm BitSeq [44] within its identification pipeline to calculate relative abundances, after aligning reads to a reference database with Bowtie 2. Unlike other alignment methods, StrainFinder[46] calculates abundances for all the genomes in the reference database using SNP frequencies after aligning reads with BWA. Because StrainFinder uses the Expectation Maximization algorithm to estimate strain frequencies, the user needs to input the expected number of strains expected to be in the sample, to ensure the best likelihood. This not only makes StrainFinder exceptionally computationally intensive, but also makes it less suitable for broad metagenomic studies with unknown number of strains.

While alignment-based detection methods work well for species with clear and well-separated sub-lineages, the selection of genomes and choice of size for the reference database is critical for applications to more closely-related strains. Some tools aim to draw on large and comprehensive databases in order to gain higher resolution. Sigma and MetaMaps offer users the opportunity to define their own reference databases and claims support for up to tens of thousands of genomes. The entirety of RefSeq (2266 genomes at time of publication) has been used as the reference database for Sigma. PathoScope generates a reference database from all genome sequences in NCBI for a given query taxID. The resulting redundancy from using a taxID which could potentially include very closely related strains, instead of a database of filtered genomes such as RefSeq, ensures coverage at all genomic levels, but can result in nonspecific

strain identification calls. Even if similar sequences are excluded, it is often not practical to have a reference genome for every genetically distinct, closely-related strain in a species. While a large reference database can increase coverage of intra-species diversity, it also requires a larger computational search space for matching reads. In addition, differentiating between closely related strains in a highly comprehensive reference database is nearly impossible and can result in an inflated number of false positive predictions. Removal of closely-related reference genomes when using BIB improves accuracy and reduces non-specific predictions to multiple unrelated strains. Therefore, proper pruning of representative reference sequences to an appropriate level of resolution is essential.

A major drawback of reference-based methods is that detection of totally novel pathogens is not possible. In contrast, assembly based methods, which reconstruct genetically distinct genotypes without need for a reference, can detect and reconstruct novel strains. When confronted with a novel strain that is not represented in the reference database, a good alignment-based detection method should output the nearest possible strain as well as the uncertainty of the match. Ultimately, meaningful results are limited to the identification of strains with reasonably close matches within the database.

Another caveat of alignment based methods is that they are dependent on details of the underlying alignment tool and its parameters. Different alignment methodologies can result in discordant results between methods and impacts our ability to perform comparisons between tools. For example, most alignment based methods use a short-read aligner[37,41,47], while DiTASiC [48] uses the pseudo alignment approach found in Kallisto [49] used for aligning RNA seq reads. Some strain identifiers (MetaMaps, Pathoscope, and MEGAN [50]) make predictions using the quality score of the alignment of each read. Sigma and BIB use Bowtie 2 as an aligner by default which reports all reads that map in multiple locations while Pathoscope and DiTASiC [48] post process multi-mapping reads within their algorithm, and StrainFinder uses BWA which randomly assigns multi-mapping reads to a specific location. Sigma additionally allows users to select their own aligner. The differences between alignment methods and their impact on results have been reported before in literature [51]. Because these strain classification methods depend on the information given via the alignment, variation at the alignment stage may have consequences throughout the entire method. Each approach can limit the ability to correctly identify strains in a sequencing set in different circumstances. The impact of these variations will ultimately depend on the species under examination and the parameters of the alignment method and how the classification methods employ the alignment information.

6.3.3 Pattern based approaches

Pattern based methods (Table 6.1) were developed to offer decreased compute time and memory requirements, as compared to assembly and alignment based methods, by focusing on a subset of the genome. These methods classify genetic diversity within a sample using a database of predefined markers, such as unique genes, SNPs, genome-specific k-mers, or fluctuations in GC content. The choice of marker type can vary based on the species, data type, and classification goals. Similar to alignment based methods, pattern based identification methods require a

reference database with which to 'learn' parameters for their statistical models. However, pattern based methods first preprocess the reference database, extract useful features, and apply these features for a new classifier algorithm, resulting in decreased run times. New sequencing reads can then be classified based off the constructed model.

An example of a method that uses a database of universal single-copy gene families as the predefined marker set is MIDAS, which aims to provide both species and strain-level taxonomic identification. MIDAS first determines species content by aligning reads to a single-copy gene database containing a single representative genome per species [52]. In order to determine strain-level information, reads are mapped to a pan-genome database containing genes from the species found in the first alignment step. Abundance estimation per strain is calculated by normalizing by the coverage of universal single copy gene families. However, this sort of strain level inference using variation in genes alone is not practical for discrimination purposes, because genes represent a smaller portion of the genome and are more likely to be conserved between strains of species[53,54]. MIDAS requires at least 1× coverage per strain to determine the presence or absence of a gene.

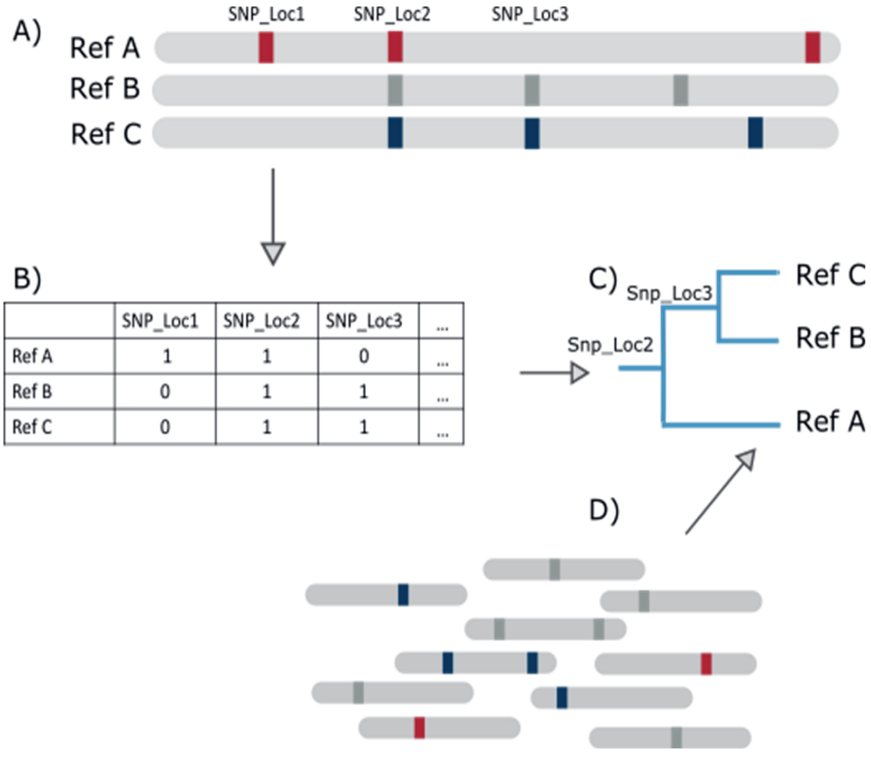
K-mers are often used in pattern based methods because unlike genes, they are sampled across the whole genome, including regions that are not especially conserved. In order to gain greater resolution than can be obtained by using only genes, GSMer identifies strains by capitalizing upon a strain-specific database of strain-specific k-mers, or GSMs (genome specific markers) [55]. Each strain in the database is represented by a set of at least 50 GSMs (optimized for k-mer size and number). If a strain has fewer than 50 unique GSMs, it is not included in the database. A strain is only identified in a read set if a perfect match for all 50 GSMs of that strain is identified within the read set, resulting in a high false negative rate and an inability to identify strains not similar to those in the database. This may work well for slow evolving and well conserved organisms that will not change and can be expected to always include the set of 50 GSMers required to be identified. But not in settings where strains are diverse and quickly changing as there is a higher chance for the set 50 GSMers required to be present to have been mutated or changed due to evolutionary drift.

Phylogenetic trees complement pattern based methods by offering a more informative database structure where paths can be indexed with a series of markers leading to a presence of a particular strain. Trees also provide an intuitive visualization of the phylogenetic placement of a strain. Given the tree, these tools map k-mers or SNPs from unknown samples onto nodes within the tree to determine phylogenetic "paths", sequences of nodes, which represent presence of a particular strain in the sample. Strain abundances are calculated based on the SNP or k-mer coverage.

SNP based tree methods differ in their SNP calling, variant filtering, tree construction, and path determination techniques. Relying solely on SNPs limits the inclusion of other types of genomic variation such as indels, which could be picked up in a k-mer based method. SNP/phylogenetic hybrid methods are particularly suitable for species with low genomic divergence like *Mycobacterium tuberculosis,* because it is a clonal organism with strains differing by very few SNPs. Gan et al, and Sahl et al (WG-FAST) have both developed tree based classification methods constructed using SNP variations between reference genomes (Figure 6.3) [56,57].

Another SNP based method, StrainEST [58], is not based on a phylogenetic tree model but instead predicts which genomes in a reference database best accounts for the frequency variation of SNPs using a linear model (LASSO).

In contrast, k-mer based tree approaches can be more suitable for species that have larger degree of genetic variation or bigger structural variations that are not detectable by only considering SNPs. They would be less efficient at differentiating strains which are only a few SNPs apart as the impacts of a genetic sequencing error are more pronounced in the tree construction and classification process when working with k-mers. Roosaare et al. (StrainSeeker [59]*)* have developed guide-tree based classification methods based on k-mers. Creation of a tree for SNP-based methods requires a collection of annotated SNP profiles for each genome. Using the SNP profiles, a new tree database is created. In the case of StrainSeeker, a phylogenetic tree detailing the relationship between reference genomes must first be provided by the user.



	SNP_Loc1	SNP_Loc2	SNP_Loc3	...
Ref A	1	1	0	...
Ref B	0	1	1	...
Ref C	0	1	1	...

Figure 6.3 **Tree Based Method Overview**. (a) Example database of genomes with SNPs present as markers. b) Representation of genome database, where 1 denotes a SNP and 0 absence of a SNP (c) SNP tree constructed based on SNPs from the database. (d) SNPs present in new reads can be matched against the tree to infer likely reference genome of origin by identifying sequences of successfully matching nodes (a path).

Another kind of approach, GOTTCHA, generates a database of unique signatures for each genome at different taxonomic levels [60]. The unique signatures of a strain are the collection of all subsequences not found in any other available sequences at the desired taxonomic level. The unique signature of an unknown query sample can then be mapped against this database to determine coverage statistics for the query's unique signature. The abundance of predicted strains is obtained through a statistic comparing the total number of mapped bases to the signature for the reference, and the number of unique bases mapped. StrainPhlAn [61] also uses species specific marker sets to classify strains, but only identifies the most abundant strain for each detected species in a metagenomic sample. The presence of other strains is assessed by calculating the number of polymorphic positions per species.

Other pattern based methods employ clustering to help delineate strains and augment pattern based detection techniques. For example, ConStrains assimilates elements of *de novo* assembly to detect genetically distinct strains [62]. Reads for each species are first mapped against species-specific marker genes using MetaPhlAn2 [63] to generate a multiple alignment, and SNPs are determined using *Samtools* [64] based on sufficient coverage criteria. The resulting SNP profiles are clustered into groups representing genetically distinct strains, with abundances calculated using a Monte-Carlo algorithm. In order to delineate strains, ConStrains requires a relatively high coverage (10×).

6.3.4 Reference database free approaches

The methods described above all depend on either the presence of genome sequences in a reference database, or the reconstruction of a genome from reads. The approaches in this section do not use a reference genome, but rather model within-sample diversity using a statistical model in order to delineate genetically distinct strains. These reference free approaches apply statistics directly from elements acquired from the sequencing read set such as SNPs or k-mers. These approaches do not attempt to identify the presence of a specific, previously-sequenced strain; rather, they utilize allele discrepancies within a WGS read set to quantify the number and proportion of unique strains present in a sample. These methods are, therefore unable to offer insight on the relationship of strains in the sample compared to previously documented strains, since there is no mapping of the sample to a database of previously seen strains. These methods only inform on the number and relative abundance of detected strains using allelic variation present in the sample. This would be useful for determining multiplicity of infection for yet uncharacterized species without many reference genomes, particularly for species present in extreme environments such as hot springs, acidic wastes, nuclear areas, uncultured soil bacteria and other under sampled locations of the world.

For example, Eyre *et al*. applied a probabilistic model to allele frequencies at specific variable sites with the underlying assumption that the sample was a mixture of two haplotypes [9]. Variable sites were defined across the whole genome as locations with ambiguous calls. As this approach is limited to modelling a maximum of two strains in the data, other methods have extended this approach to allow for the presence of multiple strains in the sample data, including *estMOI,* DEploid, and pfmix. [65–67]. Both DEploid and *estMOI* use variant calls to infer the number of

haplotypes in the dataset first locally (short regions), then globally. DEploid goes further by using a reference panel of known genomes to create a prior in their Bayesian approach to estimate the relative abundance, number of haplotypes, and their allelic states. Pfmix similarly uses a Bayesian model but do not estimate haplotypes, instead using a single reference to provide variants and allele frequencies to directly infer the number and proportions of strains from allele frequencies. Unlike reference based approaches, these methods do not output strain identities but instead the number and relative abundances of detected strains.

Table 6.1 **Tool benchmark and technical details**

Author	Method Name	Type[1]	Technical Details[2]	Sample Benchmarks[3]	Test metrics[4]	Required coverage level per strain[5]
Pulido-Tamayo, Sergio [29]	EVORha	assembly based	- java	- *E. coli* time series (lab grown) - *C. difficile* mixed infection samples	reliability score, mean absolute error, rmse	50× coverage
Quince, Christopher [34]	DESMAN	assembly based	- git/python - linear runtime - 5 strains in 117 minutes	- fecal metagenome samples - community of 100 species and 210 strains with 96 samples (synthetic)	accuracy	-
Ahn, Tae Hyuk [41]	Sigma	alignment based	- C++ - scaled for supercomputers (alignment with 10,000 cores takes 10 minutes) - sample with 5 strains takes 20 hours and 62GB RAM on a computer with 64CPU	- fecal metagenome dataset - numerous spike ins of fecal set to simulate outbreaks	accuracy, TP/FP	.02× coverage
Sankar, Aravind [47]	BIB	alignment based	- 1 million reads in 10 min on single CPU - git/python	- mixtures of 2-6 staphylococcus strains (synthetic) - *S. aureas* sample data	absolute error	
Byrd, A. L. [37,68,69]	Pathoscope	alignment based	- git/BioConda - 1 sample using 16 CPU and 256GB RAM took 17 minutes	- European *E. coli* outbreak 2011 (O104:H4) - mixed read datasets of 3 strains	TP/FP	20% genome coverage
Fischer, Martina [48]	DiTASiC	alignment based	- git/conda - requires R & python	- 3 simulated set groups - low, medium, and high complexity metagenomic benchmark datasets(synthetic) - lacks real world testing	sum of squared errors, TP/FP/FN/FP -	-
Huson, Daniel[50]	MEGAN	alignment based	- gui/java - took 180h using 64CPU for 300k reads	- Sargasso sea dataset - mammoth bone - simulation studies	FP	-

				- mostly species level testing		
Dilthey, Alexander[70]	MetaMaps	alignment based	- git/Perl - takes 16-210 hours using 262GB RAM - cannot make own DB	- simulated data - human microbiome project data (PacBio, species) - Zymo synthetic community(Oxford Nanopore Technology)	Precision, recall	-
Smillie, Christopher [46]	StrainFinder	pattern based[6]	-git/python -100 samples across 649 reference genomes using 100-200cores takes 48+hours - needs alignment file with some preprocessing as input	- 2-32 strains across 2-64 samples(synthetic) - recurrent *C. difficile* infection over time	Unifrac distance	25×
Gan, Mingyu [56]		pattern based[6]	- not available	- TB datasets	-	1× coverage
Luo, Chengwei [62]	ConStrains	pattern based[6]	- git/python - took 8.5 hours and 2 GB ram on infant gut dataset -custom DB not possible	- *E. coli* admixtures 2-7 strains (synthetic) - gut microbiome time series - microbiome time series (synthetic) - cystic fibrosis patient infection data	Jenson-Shannon divergence	10× coverage
Freitas, Tracey Allen K [60]	GOTTCHA	pattern based[6]	-git/Perl - used 16cores and 132GB RAM while being 2-5x slower than other tools -custom DB not possible	- human microbiome project mixtures of 22 genomes - spiked air filter metagenome spiked - spiked human stool - synthetic communities of 25-300 genomes	precision, recall, F-score, false discovery rate and accuracy	-
Sahl, Jason W [57]	WG-FAST	pattern based[6]	- conda - uses phylogeny	- fecal specimens *E. coli* O104:H4 outbreak	accuracy	1×
Roosaare, Märt [59]	StrainSeeker	pattern based[6]	- online web tool - Perl/R - needs 300GB space to build DB - uses 1 cpu, 512GB RAM and took 1.1min for classification	- *E. coli*, *K. pneumoniae*, *E. faecilius*, *S. enterica* isolate identification (synthetic)	accuracy	<1× coverage
Albanese, David [58]	StrainEst	pattern based[6]	- git/docker/python - takes 12-25min for a 10×-100×coverage sample using 129-591MB RAM and 4 cores	- paired strains from 4 species (synthetic) - 2 HMP mock communities (21 organisms) - specific strain in skin microbiome - cross sectional *E. coli* strains in stool samples - gut microbiome time series	Matthew Correlation Coefficient, Jensen-Shannon divergence	10× coverage

Truong, Duy Tin [61]	StrainPhlAn	pattern based[6]	- git/conda	- human microbiome	accuracy	2×
Nayfach, Stephen [52]	MIDAS	pattern based[6]	- git/docker/python - on 1CPU process 5,000 reads per second using 3 GB RAM - 1.5-2hrs for typical gut metagenome	- stool metagenomes time series - marine metagenomes	(only of genes) accuracy, TP/FP	>1× coverage
Tu, Qichao [55]	GSMer	pattern based[6]	- git/Perl scripts	- diabetes patients gut microbiome - obesity associated microbiome	TP	<0.25× (100 GSMs) >0.25× (50 GSMs)
Scholz, Matthias [71]	PanPhlAn	pattern based[6]	- git/python	- *E. coli* outbreak O104:H4 - gut microbiomes - skin microbiome - oral microbiome - marine metagenomes	F1 score	1× coverage
Eyre, David W. [9]		reference free	- R script in supplements	- *C. difficile* infected patients	RMSE	-
O'Brien, John D. [65]	pfmix	reference free	- R - for a 5 strain sample takes 10 minutes on single core	- blood from malaria patients	Mean squared error	25 reads
Assefa, Samuel [66]	*estMOI*	reference free	- git/Perl - little documentation	- clinical isolates of *P. falciparum*	accuracy	30× coverage
Zhu, Sha Joe [67]	DEploid	reference free	- R package - 1-6hours	- clinical isolates *of P. falciparum*	accuracy	1% abundance
Sobkowiak [72]	MixInfect	reference free	- R script/git - no documentation	- tested on TB samples	accuracy	10× coverage

[1]Category of algorithm
[2]Details about the computational parameters of the tool in terms of code base/runtime/memory usage/availability
[3]Example datasets tool was tested on in paper
[4]Metrics by which each method was evaluated
[5]The required coverage for the tool per stain to perform. If no value is indicated, this indicates the particular value could not be determined from the article where the method was published.
[6]Pattern based methods use a database of predefined markers to classify genetic diversity within a sample.

Table 6.2 Tool use cases and detection details

Method name	Taxonomic level[1]	A[2]	Sample setting[3]	Use cases[4]
EVORhA	strain	Y	- high coverage data	- reconstruct evolutionary trajectories - clonal populations - resolve genomes in metagenomic communities
DESMAN	strain	Y	- better with low complexity (<20 strains) communities	- environmental populations - metagenomic communities
Sigma	strain, species	Y	- made specifically to provide useful information for outbreaks	- metagenomic bio surveillance for outbreaks

BIB	strain	Y	- species with clear population structure and well-separated lineages - unsuitable for species with frequent recombination (maybe the case for many alignment methods)	- clinical use, mixed samples - flagging contaminated/problematic samples
Pathoscope	multiple levels	Y	- designed to be complete framework to analyze metagenomic data	- environmental samples - clinical samples
DiTASiC	strain	Y	- comparing abundances across samples	- general strain identification and abundance - allows for differential abundance testing across samples
MEGAN	strain, species	Y	- broad taxonomic classification	-environmental populations
MetaMaps	strain, species	Y	- long read data	- medium complexity environmental communities - medium complexity
StrainFinder	strain, species	Y	- track strain genotypes over time - specifically made to understand real world clinical problem - requires prior knowledge for number of strains	- clinical/pathogen identification - human microbiome
Gan, Mingyu	strain	Y	- specifically for TB	- clinical TB samples - mixed infections of few strains
ConStrains	strain, species	Y	- only needs one genome per species - robust against unknown strains	- clinical microbiome sets - time series data - finding specific strains within population at low abundance
GOTTCHA	user defined	Y	- designed to find low abundance populations	- clinical diagnosis - bio surveillance - community profiling
WG-FAST	strain	N	- isolate identification (single isolate & complex samples - designed for low coverage strains	- disease outbreaks - pathogen identification
StrainSeeker	strain, species	Y	- phylogeny based - identifying clade of novel strain - unable to differentiate strains with few SNV	- pathogen identification
StrainEst	strain	Y	- identifying strains of particular species - best at lower than species level - limited for poorly characterized species	- ecological/environmental samples - human/skin microbiome - molecular epidemiology
StrainPhlAn	strain, species	N	- identifies most abundant strain of particular species within metagenomes not all strains - reconstruction of stain level phylogenies of species	- human microbiome
MIDAS	strain, species	N	- cannot quantify novel species	- transmission gut microbiome
GSMer	strain, species	Y	- identify species/strain specific for well-studied organisms - possible false negatives if not all GSMs covered - false positives due to overlapping GSMs with incorrect strains	- human microbiome
PanPhlAn	strain, species	Y	- characterization of strain level gene elements - useful for population genomics where few reference genomes exist - culture free	- outbreak epidemiology - human microbiome
Eyre,David W.	strain	Y	- mixed infection detection - assumes only mixes of 2 strains	- mixed infection screening in outbreak surveillance

pfmix	strain	Y	- mixed infection detection - specifically for *P. falciparum*	- pathogen identification
estMOI	strain	N	- specifically made for *P. falciparum* - estimates multiplicity of infection - might not be possible for highly related genomes	- pathogen identification - transmission intensity
DEploid	strain	Y	- estimating mixed infections - originally developed for *P. falciparum* - can be used for any mixture of strains within species	- pathogen identification
MixInfect	strain	Y	- detecting mixed infections in TB - not suitable for non-TB species	- pathogen identification

[1]Taxonomy levels the method claims to be able to accurate identify
[2]Denotes whether a method gives the abundance of a strain.
[3]Specifics about which context the tool was originally demonstrated for
[4]Different use case scenarios that the tool can be used for or has been tested for

6.4 Comparative discussion of different methodologies

The methods mentioned in this review all aim to utilize the discriminative capability of WGS data to taxonomically classify samples at the level of individual strains within a species. These algorithms differ in required coverage, the number of strains that can be detected, the ability to detect higher level taxa (Table 6.2), and other criteria.

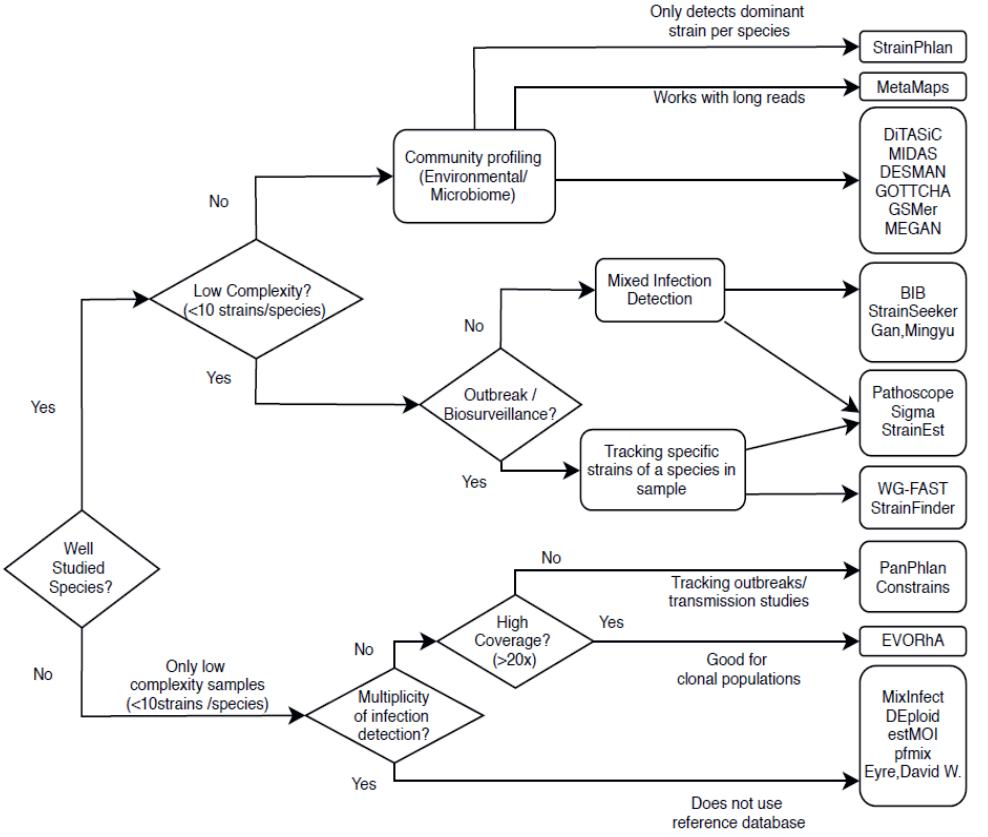


Figure 6.4 **Flow chart of tool selection depending on scenario**. Guide chart showing which tools can be used in which use case. Presence of a tool under one use case doesn't necessarily exclude it from being applicable to another use case.

To help guide tool selection we have made a flow chart (Figure 6.4) showing which types of tools would work well with different use cases. Alignment and pattern based methods are more broadly applicable because they can be used on samples with lower coverage levels of the species of interest (<1×) than assembly based methods, which can require high coverage levels (50× for EVORhA). Although assembly based methods can provide the possible genomic sequence of a strain of interest and are not limited to the breadth of the genomes present in a reference database, alignment methods are faster and can be used at lower coverages. To taxonomically classify or examine intra-species heterogeneity within an isolate culture expected to contain a single, well-studied species (such as *E. coli)*, both alignment and pattern based methods can be used, such as Pathoscope and StrainSeeker, as these methods require prior knowledge of a species. Both methods have tools that require differing coverage levels to identify strains (Table 6.1), from less than 1× in the case of StrainSeeker and Pathoscope, up to a minimum of 10× coverage in the case of ConStrains). However, if the single species present in the isolate sample

is not as well-studied, then assembly methods such as EVORhA are more suited, as they are not as dependent on prior knowledge encapsulated in a reference database. Because assembly methods require sufficient coverage (50× for EVORhA) to resolve haplotypes, these methods are not suitable for communities of samples with low coverages. Certain methods quantify multiplicity of infection or relative abundance of strains within a sample using allelic variations within the dataset and do not require a database of known genomes. These reference-free tools are useful when the relationships between strains in a single-species sample are of interest, rather than the exact strain identities or their relationships to previously studied strains. This would be suitable for testing multiplicities of strains in uncultured soil samples or other extreme environments which are still under sampled. Reference-free approaches can also be applied for well characterized species, however since pattern and alignment based tools can also offer strain identity – these might be preferred due to the extra information given.

When exploring strain level variety in the context of a metagenomic sample containing multiple species, alignment and pattern based methods such as GSMer or Sigma are able to classify at both the species and strain level. Within metagenomic samples, the strains of interest could be present at very low coverages, making assembly based methods not suitable for this task.

Phylogenetic-based methods such as those of Gan et al. [56] can also track evolutionary divergence of the same strain within a longitudinal metagenomic samples, which requires a high level of resolution to detect small changes in gene sequences. These methods work at low coverage levels, and have the advantage of including a visual representation of the underlying decision process which can be easier to explain and understand. The phylogenetic framework also offers users the ability to sanity check results. For example, multiple closely related strains can be detected when the ‘true’ strain is not present in the database.

When focusing on one particular species within a metagenomic sample, methods such as Sigma and Pathoscope are able to remove DNA from other species, which increases the specificity of the desired species and reduces downstream computation time. This is useful for diagnostics which aim to detect strains from only a particular pathogenic species, for example, when iteratively looking for strains within a particular species in a metagenomic sample.

Though this might provide a slight boost in computational efficiency, Sigma and Pathoscope are still both computationally intensive programs.

Ease of use and speed of analysis are both important concerns when considering a metagenomic tool. Table 6.2 details the different machine requirements and speed tests given by the methods reported in this paper. Though versatile and adaptable to different scenarios, tools requiring extensive mapping to a reference database can be extremely computationally intensive. Sigma required nearly 20 hours resolving a single 5 strain community (20 million reads) against a database of 2,266 reference genomes with 62GB of memory and 64 cores. StrainFinder, another alignment method, took more than 48 hours with 100-200 cores for 100 samples. Some methods were tested in high performance computing environments (i.e. Pathoscope, MEGAN, MetaMaps, GOTTCHA, all > 100GB memory) which may not always be available for clinicians. Additionally, tools requiring a database typically only report times/requirements to process a sample, but rarely include the time required to generate a custom database. We were only able to find both values

for StrainSeeker, which process samples relatively quickly (<2 minutes) but suggests 300GB of space and 512GB of ram available to generate a database. In terms of usability, almost all of the tools were made to run in a Linux environment, therefore requiring some level of computational expertise in order to navigate requirements and installation setups. Few tools offer an online accessible functionality (MEGAN and StrainSeeker). That being said, certain tools are bundled in easy to install package managers like Conda and R (i.e. DEploid, pfmix, StrainPhlAn), while others only offer a collection of scripts (i.e. MixInfect, and Eyre et al). Due to the requirements for installation and use (Bash/Linux), using most of these methods would require some bioinformatics knowledge. Further work would need to go into making these tools accessible and open for general use, such as online web tools, or a easy to use/install gui.

Most of the methods described in this review have not been benchmarked across all possible use case scenarios in a systematic or independent manner; therefore, a researcher using these tools will need to carefully determine whether a particular tool would work for their data type of interest. We discuss more about benchmarking in the next section.

6.5 Method evaluation, benchmarking and simulation

Thorough and robust benchmarking of algorithms for a particular application and data type is critical. As this field is relatively new, there has yet to be a proper comparative study benchmarking the efficiency, accuracy, and specificity of these methods in a diversity of application domains: clinical pathogens [73,74], microbiomes [75,76] and industrial biotechnology [42,77,78] as examples.

The types of validation that have been performed for each method are indicated in Table 6.1. For all tools1, an initial validation of model performance was performed using *in silico* simulated reads of known composition, generated from genomes of known host strains using tools such as *MetaSim, Grinder,* and *Art* [79–81]. Alternatively, sequencing reads from presumed pure strains can be used. Testing applicability to strain mixes involves constructing a more complex synthetic dataset containing a mixture of varying quantities of individual strain read sets. Factors that must be considered in the construction of synthetic validation datasets include: 1) Determining the actual sequencing depth necessary to be able to identify a particular strain in a read set and number of reads to use. 2) The diversity in strain composition in terms of taxonomic levels that should be represented or background non-target species 3) The level of complexity that needs to be introduced in the reads (in terms of SNVs and genomic distance between strains) and 4) the scalability of the method to fluctuation in sample size (e.g. low abundance strains in large sample sets). Validation on synthetic datasets addresses performance of the algorithms in the best-case scenario. Subsequent to these validation experiments, performance needs to be examined on test-case ‘real’ samples, as this is often presents a much greater challenge than testing on in-silico-generated datasets.

In order to compare the results of benchmarking different tools, metrics for comparing results across different types of outputs from various tools must be carefully chosen. The published benchmarking methods for the tools described in Table 6.1 use a variety of different metrics. The

most common method employed for the published tools involves testing the specific algorithm on a dataset of known diversity and abundance, and comparing accuracy metrics. For alignment- and pattern based methods, a true and false positive would be defined as whether the algorithm was able to detect the correct strain within the sample, or whether it detected the wrong strain, respectively. A false negative would be defined if the algorithm failed to detect a strain present in the sample, and a true negative would be called if the algorithm did not output any strains not present. An important consideration in the assessment of true negatives is whether the algorithm informs the user of the uncertainty of the match and outputs the nearest strain. Most methods mentioned in this paper quantified the reliability of their method by either calculating the true positive rate/false discovery rate or by checking manually whether the results were correct.

In addition to simply identifying which strains are present or absent in a sample, additional metrics must assess the accuracy in estimating strain abundances. One method to do this, used by the assembly based detection method, EVORhA, uses the mean absolute error (MAE) metric between the true abundances and estimated abundances. In addition, they also calculated the root mean squared error (RMSE), which was also used by Eyre *et al*. Another method to assess accuracy in strain abundance is the Jenson-Shannon divergence, which was used in ConStrains to measure their prediction accuracy.

A comprehensive comparison and benchmarking of these tools is needed to provide further insight into the efficiency of these tools at performing strain-level identification on a wide range of sample types, be it metagenomic, clinical, or cultures. This benchmarking strategy would need to deal with the nuances between tools, as they have different goals, different use-case scenarios, and different criteria for success. It might be possible to conduct these comprehensive benchmarks in categories such that similar tools could be evaluated together on novel datasets with a common evaluation metric.

6.6 Conclusion and future directions

Whole genome sequencing of microbial populations has the capability to offer a view into genetic diversity at varying taxonomic levels. Current widely-used taxonomic classifiers allow for the identification of species within WGS sets. However, algorithms for finer-grained classification, at the individual strain level within a species, are still relatively new. Such techniques have the capacity to greatly impact healthcare and other fields by precise tracking of disease outbreaks, differentiation of commensal and pathogenic strains, and linking strain level genotypic traits with phenotypic characteristics of clinical and industrial importance [42,73–78].One assumption almost universally made within taxonomic tools is that a direct relationship exists between strain read coverage and strain abundance in the sample. As such, calculations of strain abundance levels take into account the variations of coverage across variant sites or reads. Though intuitive, none of the tools presented here presented analysis to prove this assumption. Conducting such verification steps is particularly important for tools focusing on clinical use and pathogen identification, where it is typical for a culturing step to be conducted before sequencing. In actuality, there could be many reasons why read abundance does not directly reflect the

composition of the sample: isolation technique (culture sweep vs single colony isolation), contamination skewing read depth, or the sequencing process itself.

There are numerous ways in which current strain identification methods can improve their benchmarking. Firstly, very few algorithms tested the performance of their tools on multiple (>2) low abundance strains (<1-2x). Detecting low abundance strains would be preferred for microbial communities such as the gut, where specific strains exhibit differing pathogenicity. Secondly, no methods quantified or benchmarked how genetically distant a strain needs to be in order to properly delineate it. Third, there are no tools that allow a user to compare strains within and across samples, which would be useful for transmission studies. Lastly, delineating extremely closely related strains remains a difficult problem for the metagenomic tools. Many tools requiring a reference database remove genomes from the database that are extremely close together or self-report that they would not work well with highly related genomes [47,58,66]. Such analysis remains difficult due to the problems that arise when considering closely related strains such as an increase in false positives due to both strains being reported when only one is actually present or problems within the model itself driven by high levels of collinearity. The difficulty with detecting extremely close strains is further compounded due to the ambiguous definition of a strain.

The methods detailed in this literature review are almost all directed towards sequencing technologies that produce reads from mixtures of cells. Direct sequencing of individual cells would bypass this need to computationally subdivide reads produced from current NGS technologies into those originating from different strains. Single-cell sequencing strategies such as Drop-Seq [82] and 10X Genomics [83] are rapidly improving to provide a systematic and comprehensive view of the genetic diversity of complex communities. Having sequencing data originating from individual cells would greatly simplify studies of heterogeneous populations of strains. However, there are still technical difficulties to overcome before single-cell sequencing becomes widely adopted. It is probable that the next iteration of strain-level identification algorithms will be focused on such technologies. One pioneering example is *MetaSort,* which combines the advantages of both WGS and single cell sequencing data [84]. This method assembles genomes from both WGS reads and single cell sequencing reads and integrates the two using a machine-learning algorithm, resulting in genomes present in the sample. The increased resolution from single cell sequencing based detection is likely to uncover novel forms of genetic heterogeneity. In addition, advances in long read sequencing continue to change the scope and direction of strain-level detection in metagenomic samples. Longer read lengths could make it easier and more practical to phase haplotypes, as well as identify strains with fewer reads. We have touched upon a few long-read methods (MetaMaps and that of Somervile et al), and still expect that many more will be released as this platform of sequencing continues to gain in popularity.

The ability to quantify and detect bacterial strains within heterogeneous environments has applications in numerous fields including diagnostics[85], clinical studies for the microbiome [86], bio surveillance[41], tracking transmission of infectious strains in an outbreak [37,41,52], providing insight into the spread of antibiotic resistance [87], tracking progression of within-host bacterial evolution[29] and exploring diverse environments [88]. We look forward to the wide range of applications and effects these tools will have in shaping and progressing sequencing based research.

6.7 References

1. Fournier P-E, Dubourg G, Raoult D. Clinical detection and characterization of bacterial pathogens in the genomics era. Genome Med. 2014; 6:114

2. Maxson T, Mitchell DA. Targeted treatment for bacterial infections: prospects for pathogen-specific antibiotics coupled with rapid diagnostics. Tetrahedron 2016; 72:3609–3624

3. Balmer O, Tanner M. Prevalence and implications of multiple-strain infections. Lancet Infect. Dis. 2011; 11:868–878

4. Alizon S, de Roode JC, Michalakis Y. Multiple infections and the evolution of virulence. Ecol. Lett. 2013; 16:556–567

5. Deurenberg RH, Bathoorn E, Chlebowicz MA, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J. Biotechnol. 2017; 243:16–24

6. Cohen T, van Helden PD, Wilson D, et al. Mixed-strain Mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. Clin. Microbiol. Rev. 2012; 25:708–719

7. Marshall JA. Mixed Infections of Intestinal Viruses and Bacteria in Humans. Polymicrob. Dis. 2002;

8. Plazzotta G, Cohen T, Colijn C. Magnitude and sources of bias in the detection of mixed strain M. tuberculosis infection. J. Theor. Biol. 2015; 368:67–73

9. Eyre DW, Cule ML, Griffiths D, et al. Detection of Mixed Infection from Bacterial Whole Genome Sequence Data Allows Assessment of Its Role in Clostridium difficile Transmission. PLoS Comput. Biol. 2013; 9:

10. Eyre DW, Walker AS, Griffiths D, et al. Clostridium difficile mixed infection and reinfection. J. Clin. Microbiol. 2012; 50:142–144

11. Esposito S, Bosis S, Cavagna R, et al. Characteristics of Streptococcus pneumoniae and atypical bacterial infections in children 2-5 years of age with community-acquired pneumonia. Clin. Infect. Dis. 2002; 35:1345–52

12. Minagawa S, Takayanagi N, Hara K, et al. [Clinical features of mixed infections in patients with Streptococcus pneumoniae pneumonia]. Nihon Kokyuki Gakkai Zasshi 2008; 46:278–284

13. Navarro Y, Herranz M, Pérez-Lago L, et al. Systematic survey of clonal complexity in tuberculosis at a populational level and detailed characterization of the isolates involved. J. Clin. Microbiol. 2011; 49:4131–4137

14. Huang HY, Tsai YS, Lee JJ, et al. Mixed infection with Beijing and non-Beijing strains and drug resistance pattern of Mycobacterium tuberculosis. J. Clin. Microbiol. 2010; 48:4474–4480

15. Cespedes C, Said-Salim B, Miller M, et al. The clonality of Staphylococcus aureus nasal carriage. J Infect Dis 2005; 191:444–452

16. Lessing MP, Jordens JZ, Bowler IC. Molecular epidemiology of a multiple strain outbreak of methicillin-resistant Staphylococcus aureus amongst patients and staff. J. Hosp. Infect. 1995; 31:253–60

17. El-Halfawy OM, Valvano MA. Antimicrobial heteroresistance: An emerging field in need of

clarity. Clin. Microbiol. Rev. 2015; 28:191–207

18. Falagas ME, Makris GC, Dimopoulos G, et al. Heteroresistance: A concern of increasing clinical significance? Clin. Microbiol. Infect. 2008; 14:101–104

19. Frank SA. Models of Parasite Virulence. Q. Rev. Biol. 1996; 71:37–78

20. Wood DE, Salzberg SLS, Venter C, et al. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15:R46

21. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods 2015; 12:902–903

22. Mande SS, Mohammed MH, Ghosh TS. Classification of metagenomic sequences: Methods and challenges. Brief. Bioinform. 2012; 13:669–681

23. Teeling H, Glöckner FO. Current opportunities and challenges in microbial metagenome analysis-A bioinformatic perspective. Brief. Bioinform. 2012; 13:728–742

24. Hunter CI, Mitchell A, Jones P, et al. Metagenomic analysis: The challenge of the data bonanza. Brief. Bioinform. 2012; 13:743–746

25. Goldman D, Domschke K. Making sense of deep sequencing. Int. J. Neuropsychopharmacol. 2014; 17:1717–1725

26. Marx V. Microbiology: the road to strain-level identification. Nat. Methods 2016; 13:401–404

27. Yuan S, Cohen DB, Ravel J, et al. Evaluation of methods for the extraction and purification of DNA from the human microbiome. PLoS.ONE. 2012; 7:e33865

28. Votintseva AA, Bradley P, Pankhurst L, et al. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. J. Clin. Microbiol. 2017; 55:1285–1298

29. Pulido-Tamayo S, Sánchez-Rodríguez A, Swings T, et al. Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. Nucleic Acids Res. 2015; 43:e105

30. Bertrand D, Shaw J, Kalathiyappan M, et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat. Biotechnol. 2019; 37:937–944

31. Tsai YC, Conlan S, Deming C, et al. Resolving the complexity of human skin metagenomes using single-molecule sequencing. MBio 2016; 7:

32. Somerville V, Lutz S, Schmid M, et al. Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. BMC Microbiol. 2019; 19:

33. Kolmogorov M, Yuan J, Lin Y, et al. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 2019; 37:540–546

34. Quince C, Delmont TO, Raguideau S, et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol. 2017; 18:181

35. Zagordi O, Bhattacharya A, Eriksson N, et al. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics 2011; 12:119

36. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. 2010; 11:473–483

37. Hong C, Manimaran S, Shen Y, et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2014; 2:33

38. Clement NL, Snell Q, Clement MJ, et al. The GNUMAP algorithm: Unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. Bioinformatics 2009; 26:38–45

39. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012; 9:357–359

40. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402

41. Ahn TH, Chai J, Pan C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics 2015; 31:170–177

42. Walsh AM, Crispie F, Daari K, et al. Strain-Level Metagenomic Analysis of the Fermented Dairy Beverage Nunu Highlights Potential Food Safety Risks. Appl. Environ. Microbiol. 2017; 83:e01144-17

43. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. 2011; 12:15

44. Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics 2012; 28:1721–1728

45. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 2012; 7:562–78

46. Smillie CS, Sauk J, Gevers D, et al. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. Cell Host Microbe 2018; 23:229-240.e5

47. Sankar A, Malone B, Bayliss S, et al. Bayesian identification of bacterial strains from sequencing data. bioRxiv 2015; 1–16

48. Fischer M, Strauch B, Renard BY. Abundance estimation and differential testing on strain level in metagenomics data. Bioinformatics 2017; 33:i124–i132

49. Bray NL, Pimentel H, Melsted P, et al. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 2016; 34:525–527

50. Huson DH, Auch AF, Qi J, et al. MEGAN analysis of metagenomic data. Genome Res. 2007; 17:377–386

51. Canzar S, Salzberg SL. Short Read Mapping: An Algorithmic Tour. Proc. IEEE 2017; 105:436–458

52. Nayfach S, Rodriguez-Mueller B, Garud N, et al. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 2016; 26:1612–1625

53. Jordan IK, Rogozin IB, Wolf YI, et al. Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. Genome Res. 2002; 12:962–968

54. Martín MJ, Herrero J, Mateos A, et al. Comparing bacterial genomes through conservation profiles. Genome Res. 2003; 13:991–998

55. Tu Q, He Z, Zhou J. Strain/species identification in metagenomes using genome-specific markers. Nucleic Acids Res. 2014; 42:1–12

56. Gan M, Liu Q, Yang C, et al. Deep whole-genome sequencing to detect mixed infection of mycobacterium tuberculosis. PLoS One 2016; 11:1–14

57. Sahl JW, Schupp JM, Rasko DA, et al. Phylogenetically typing bacterial strains from partial SNP genotypes observed from direct sequencing of clinical specimen metagenomic data. Genome Med. 2015; 7:52

58. Albanese D, Donati C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat. Commun. 2017; 8:2260

59. Roosaare M, Vaher M, Kaplinski L, et al. StrainSeeker : fast identification of bacterial strains from unassembled sequencing reads using user-provided guide trees . bioRxiv 2016; 1–20

60. Freitas TAK, Li P-E, Scholz MB, et al. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res. 2015; gkv180

61. Truong DT, Tett A, Pasolli E, et al. Microbial strain-level population structure & genetic diversity from metagenomes. Genome Res. 2017;

62. Luo C, Knight R, Siljander H, et al. ConStrains identifies microbial strains in metagenomic datasets. Nat. Biotechnol. 2015; 33:0–4

63. Segata N, Waldron L, Ballarini A, et al. Metagenomic microbial community profiling using unique clade- specific marker genes. Nat Methods 2013; 9:811–814

64. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011; 27:2987–2993

65. O'Brien JD, Iqbal Z, Wendler J, et al. Inferring Strain Mixture within Clinical Plasmodium falciparum Isolates from Genomic Sequence Data. PLoS Comput. Biol. 2016; 12:1–20

66. Assefa SA, Preston MD, Campino S, et al. EstMOI: Estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics 2014; 30:1292–1294

67. Zhu SJ, Almagro-garcia J, Mcvean G. Deconvoluting multiple infections in Plasmodium falciparum from high throughput sequencing data. 2017; 1–13

68. Byrd AL, Perez-Rogers JF, Manimaran S, et al. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC Bioinformatics 2014; 15:262

69. Francis OE, Bendall M, Manimaran S, et al. Pathoscope : Species identification and strain attribution with unassembled sequencing data Pathoscope : Species identification and strain attribution with unassembled sequencing data. Genome Res. 2013; 1721–1729

70. Dilthey AT, Jain C, Koren S, et al. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. Nat. Commun. 2019; 10:

71. Scholz M, Ward D V, Pasolli E, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat. Methods 2016; 13:435–438

72. Sobkowiak B, Glynn JR, Houben RMGJ, et al. Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data. BMC Genomics 2018; 19:613

73. Cassir N, Benamar S, La Scola B. Clostridium butyricum: from beneficial to a new emerging pathogen. Clin. Microbiol. Infect. 2016; 22:37–45

74. Ward DV, Scholz M, Zolfo M, et al. Metagenomic Sequencing with Strain-Level Resolution Implicates Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants. Cell Rep. 2016; 14:2912–2924

75. Fang X, Monk JM, Nurk S, et al. Metagenomics-Based, Strain-Level Analysis of Escherichia coli From a Time-Series of Microbiome Samples From a Crohn's Disease Patient. Front. Microbiol. 2018; 9:2559

76. Goltsman DSA, Sun CL, Proctor DM, et al. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. Genome Res. 2018; 28:1467–1480

77. De Filippis F, La Storia A, Villani F, et al. Strain-Level Diversity Analysis of Pseudomonas fragi after In Situ Pangenome Reconstruction Shows Distinctive Spoilage-Associated Metabolic Traits Clearly Selected by Different Storage Conditions. Appl. Environ. Microbiol. 2019; 85:e02212-18

78. Capece A, Granchi L, Guerrini S, et al. Diversity of Saccharomyces cerevisiae Strains Isolated from Two Italian Wine-Producing Regions. Front. Microbiol. 2016; 7:1018

79. Richter DC, Ott F, Auch AF, et al. MetaSim: A Sequencing Simulator for Genomics and Metagenomics. Handb. Mol. Microb. Ecol. I Metagenomics Complement. Approaches 2011; 417–421

80. Angly FE, Willner D, Rohwer F, et al. Grinder: A versatile amplicon and shotgun sequence simulator. Nucleic Acids Res. 2012; 40:

81. Huang W, Li L, Myers JR, et al. ART: A next-generation sequencing read simulator. Bioinformatics 2012; 28:593–594

82. Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015; 161:1202–1214

83. Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat. Commun. 2017; 8:

84. Ji P, Zhang Y, Wang J, et al. MetaSort untangles metagenome assembly by reducing microbial community complexity. Nat. Commun. 2017; 8:

85. Dekkera JP. Metagenomics for clinical infectious disease diagnostics steps closer to reality. J. Clin. Microbiol. 2018; 56:

86. Wang WL, Xu SY, Ren ZG, et al. Application of metagenomics in the human gut microbiome. World J. Gastroenterol. 2015; 21:803–814

87. Sukhum K V., Diorio-Toth L, Dantas G. Genomic and Metagenomic Approaches for Predictive Surveillance of Emerging Pathogens and Antibiotic Resistance. Clin. Pharmacol. Ther. 2019; 106:512–524

88. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat. Rev. Genet. 2005; 6:805–814

7 Discussion

Research into the bacteria, *Mycobacterium tuberculosis,* has been revolutionized by advances in whole genome sequencing (WGS) technologies. An active challenge within the research community is finding innovative ways to process and extract information from the data produced by these technologies. This requires the creation of bioinformatics tools and applications specifically focused on the aims of TB research, with bridging the gap between research and clinical practice as the ultimate goal. Within the four years this thesis was written, tools, studies, and research focusing on WGS based analysis of TB has exploded. Paired with the increase of TB WGS research, were the advances in sequencing and molecular techniques. For example: the great reduction in long-read sequencing error rates, new protocols for metagenomics library preparation, the upgrade of the WHO endorsed TB diagnostic GeneXPert, amongst others. While there is much hope and expectations in the power for WGS to quell the TB burden, still many problems remain for clinical adoption. This thesis has presented several applications for WGS guided analysis of TB research and now, in this final chapter, I will share my thoughts on active problems for the TB research community and possible directions for the future.

TB research focused on diagnostics, outbreak reconstruction, MDR and XDR surveillance, and reinfection studies, have typically failed to consider the impact within patient heterogeneity could have on the accuracy of results or downstream clinical effects. This is particularly an issue as different strains of MTB can have different phenotypes, and when left undetected, this heterogeneity can complicate research and confound treatment. There needs to be better characterization of individual strains of TB and how they relate to resistance, pathogenicity, and immune response. In chapter 2, we developed an algorithm to tackle this problem. The algorithm detects and identifies individual strains of MTB within WGS data of cultured TB isolates. It additionally outputs antibiotic resistance mutations present in the isolates. This tool was innovative compared to other similar metagenomics tools in that it focused specifically on MTB, which has lower variation between strains compared to other species. We showed the tool could properly discriminate strains and had higher discriminative power compared to currently available tools. With this tool, we provide the greater MTB field another resource to gain insights into their WGS data.

Our algorithm suffers from similar pitfalls to other innovative tools within the TB field. It is a command line tool written and tested within a Linux environment. This is the case for many bioinformatics tools for TB and is a problem for researchers or clinicians who are not so tech savvy. Specific problems include the fact these tools can be 1) quite complicated to compile and install on a computing device 2) not tested with different kinds of data/computing environments 3) dependent on funding and/or an active contributor, the lack of which can lead to maintenance issues and unreliability and 4) not user friendly as many lack proper instructions and/or use difficult to understand language. Wider adoption of bioinformatics tools within the TB community and especially in the clinic, can be achieved by prioritizing user friendliness, increasing education of bioinformatics amongst clinicians/microbiologists, and including provisions for tool maintenance within funding pipelines.

Our algorithm was originally intended to be used for diagnostic purposes. Indeed, the dream for many would be to replace the current diagnostic procedure involving cultured sputum with a complete suite of rapid WGS based tools. However, nearly all WGS based analysis requires a culture step in order to extract the sufficient quantities of MTB genetic material. A rapid WGS based diagnostic would only be possible with direct to sample sequencing. Culture free WGS experiments directly from sputum would have numerous benefits. It would remove the time spent in culture, but an additional often overlooked benefit is the significant reduction of the bias inherent in the culturing process. Within host heterogeneity of MTB is well documented, however through culture, be it a plate sweep or a single colony isolate, some of this heterogeneity is lost [1]. In studies it was shown that clonal composition changed dramatically after the culture step, such that in a subset of cases only a single strain remained [2,3]. In actuality, the true impact of the culture step on the ability to detect minority strains is still not very well studied [4]. Therefore clinicians and researchers are left with merely a glimpse of the underlying bacterial population within a patient. Sequencing directly from a patient sample would not completely solve this bias, but this would be a big step in its reduction [5]. Efforts have been made to bring culture free WGS to fruition [6–8], but issues pertaining to human/microbial contamination and the acquisition of sufficient sequencing depth have limited progress. Continued research into improving the library prep stage and downstream bioinformatics filtering techniques could enable direct to sample sequencing of TB sometime in the not too distant future.

Chapters 2-4 all utilize short-read sequencing technologies. Although short-reads are the dominant mode of sequencing, new technologies are steadily under development. Sequencing technologies by Oxford Nanopore Technologies (ONT) and PacBio enable the coverage of thousands of base pairs within a single read. ONT comes with the added benefit of being highly portable due to its small size. The growing adoption of long-read sequencing platforms will spawn the creation of new bioinformatics tools and pipelines. It would behove already existing bioinformatics tools, such as the one presented in this thesis, to reflect and adapt to the new technologies.

Analysis conducted in chapters 2-4 all discarded variation within the PE/PPE region, a phenomenon that was tackled in Chapter 5. Indeed, a big advantage for WGS over molecular techniques is in its ability to capture variation across the whole genome. Despite this capability, a large proportion of the TB genome is discarded during analysis – the PE/PPE gene family which constitute 10% of the genome. This gene family hosts much of the variation within TB and is replete with repetitive GC-rich regions. Because bioinformatics algorithms struggle with processing short-read data from this region, this information is typically ignored, despite the potentially informative variation present within these regions. In chapter 5 we showed that this need not be the case for more than 80% of the PE/PPE genes. We used high quality nanopore assemblies to shed light in the genetic variation within these genes. We found that nearly all PE genes are quite conserved across lineages. And despite being highly variable, we observed considerable conservation in several PPE/PE-PGRS genes. We also showed that mappability of these genes to the H37Rv reference genome was not as error prone as previously estimated, suggesting that they do not need to be discarded in analysis. Therefore, our research provided the TB community a much needed sanction on the use of these PE/PPE genes within bioinformatics analysis.

A shortcoming present within chapters 2-4 is in the way I predict antibiotic resistance within the MTB isolates. In these chapters, the same list of SNPs were used to detect an antibiotic resistant genotype within an isolate. The list used is by no means the most exhaustive list of resistance conferring mutations in TB, nor is it the only list available. The latter point cannot be overstated. There are tens of studies, reviews, meta-analysis and software devoted to the accumulation and prediction of MTB antibiotic resistance [9–21]. Major attempts at standardizing these lists into a widely used database have been put forward but some of these databases have not been maintained, and despite the appearances of new software/databases, long term maintenance and accessibility still remains an active challenge within the TB community [22]. Continued integration of available databases/lists are necessary and new studies should aim to associate their list curation/research with respect to the major antibiotic resistance database present at the time. Greater amounts of communication between organizations curating, developing and applying antibiotic resistance mutation software will go a long way to expediting the integration process.

As stated, the list of SNPs we used to determine antibiotic resistance was not exhaustive, not only due to lack of consensus of mutation lists within the TB community, but also due to incomplete knowledge of resistance conferring genotypes. It is unknown how much of TB antibiotic resistance conferring variation is still unknown! This gap in our collective knowledge has led to assumptions/conclusions being made based on incomplete data. This is doubly true in the case of heteroresistance, which similar to the phenomena of mixed infections has been overlooked in WGS based analysis. Lack of detected resistance either by knowledge deficit or undetected heteroresistance could 1) mislead diagnostic efforts causing future spread of DR TB by inadequate treatment. Further WGS studies are needed to illuminate markers of MTB resistance, which can then be implemented in the creation of robust diagnostic algorithms. And 2) lead to unreproducible research as future studies arrive to different sets of conclusions given a more robust view on resistance mechanisms. For heteroresistance, we showed this was the case in Chapter 3 on a global set of MTB isolates. We revealed the extent of undetected heteroresistance across our set, a large portion of which was found within isolates containing mixtures of strains.

In Chapter 3 we also revealed the extent of mixed infections present within past research studies. We pointed out the high association of heteroresistance and mixed infections with MDR/XDR-TB. We also examined the probable impact of mixed infections on transmission studies. This all was enabled through the accumulation and curation of a large collection of publically available MTB isolates from the sequence read archive (SRA) [23]. This study could have been enriched with higher quality meta-data information, such as resistance phenotypes, dating of isolates, isolation method, all of which proved much more difficult to obtain than the sequencing data itself. This points to a general drawback of publically available data. As each research institute has their own guidelines, protocols, and methods to share data, it becomes increasingly difficult to compare results across studies. This limits the ability to conduct proper meta-analysis and/or gain new insights from previously sequenced data. Standardizing the ways data is uploaded would be beneficial for future studies. Databases containing clean genomics data paired with cleaned metadata already exist [21] whose format could be used as a model for those researchers planning on publically sharing their data.

The benefits of WGS in the clinic are undeniable. Deeper insight into pathogenic genetic material present within a patient is useful for surveillance and diagnostic purposes. However, most of TB genomic studies have been performed in low incidence countries with plentiful access to funding. These countries can afford the price tag that comes with sequencing, which despite the decreasing rates, is still not as affordable as simpler molecular techniques. It is unlikely that high burden low-income countries will have the resources available to control TB using WGS tracking/tracing systems in the same capacity as high-income countries can. However insights gained from the research by richer countries can be beneficial to the global community. Additionally, genomics can discover new markers that could be implemented in the creation of more affordable point of care devices that combine both diagnosis and DST testing.

7.1 References

1. Mallard K, McNerney R, Crampin AC, et al. Molecular detection of mixed infections of Mycobacterium tuberculosis strains in sputum samples from patients in Karonga District, Malawi. J. Clin. Microbiol. 2010;

2. Martín A, Herranz M, Ruiz Serrano MJ, et al. The clonal composition of Mycobacterium tuberculosis in clinical specimens could be modified by culture. Tuberculosis 2010;

3. Metcalfe JZ, Streicher E, Theron G, et al. Mycobacterium tuberculosis subculture results in loss of potentially clinically relevant heteroresistance. Antimicrob. Agents Chemother. 2017;

4. Cohen T, van Helden PD, Wilson D, et al. Mixed-strain Mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. Clin. Microbiol. Rev. 2012; 25:708–719

5. Nimmo C, Shaw LP, Doyle R, et al. Whole genome sequencing Mycobacterium tuberculosis directly from sputum identifies more genetic diversity than sequencing from culture. BMC Genomics 2019; 20:1–9

6. Votintseva AA, Bradley P, Pankhurst L, et al. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. J. Clin. Microbiol. 2017; 55:1285–1298

7. Doyle RM, Burgess C, Williams R, et al. Direct whole-genome sequencing of sputum accurately identifies drug-resistant mycobacterium tuberculosis faster than MGIT culture sequencing. J. Clin. Microbiol. 2018;

8. Doughty EL, Sergeant MJ, Adetifa I, et al. Culture-independent detection and characterisation of Mycobacterium tuberculosis and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer. PeerJ 2014;

9. Ngo TM, Teo YY. Genomic prediction of tuberculosis drug-resistance: Benchmarking existing databases and prediction algorithms. BMC Bioinformatics 2019;

10. Flandrois JP, Lina G, Dumitrescu O. MUBII-TB-DB: A database of mutations associated with antibiotic resistance in Mycobacterium tuberculosis. BMC Bioinformatics 2014;

11. Feuerriegel S, Schleusener V, Beckert P, et al. PhyResSE: A web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing

data. J. Clin. Microbiol. 2015;

12. Coll F, McNerney R, Preston MD, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;

13. Sekizuka T, Yamashita A, Murase Y, et al. TGS-TB: Total genotyping solution for Mycobacterium tuberculosis using short-read whole-genome sequencing. PLoS One 2015; 10:1–12

14. Sandgren A, Strong M, Muthukrishnan P, et al. Tuberculosis drug resistance mutation database. PLoS Med. 2009;

15. Nebenzahl-Guimaraes H, Jacobson KR, Farhat MR, et al. Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in Mycobacterium tuberculosis. J. Antimicrob. Chemother. 2014;

16. Laurenzo D, Mousa SA. Mechanisms of drug resistance in Mycobacterium tuberculosis and current status of rapid molecular diagnostic testing. Acta Trop. 2011;

17. Maruri F, Sterling TR, Kaiga AW, et al. A systematic review of gyrase mutations associated with fluoroquinolone-resistant Mycobacterium tuberculosis and a proposed gyrase numbering system. J. Antimicrob. Chemother. 2012;

18. Georghiou SB, Magana M, Garfein RS, et al. Evaluation of genetic mutations associated with mycobacterium tuberculosis resistance to amikacin, kanamycin and capreomycin: A systematic review. PLoS One 2012;

19. Miotto P, Tessema B, Tagliani E, et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. Eur. Respir. J. 2017;

20. . Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N. Engl. J. Med. 2018;

21. Starks AM, Aviles E, Cirillo DM, et al. Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform. Clin. Infect. Dis. 2015;

22. Cohen KA, Manson AL, Desjardins CA, et al. Deciphering drug resistance in Mycobacterium tuberculosis using whole-genome sequencing: Progress, promise, and challenges. Genome Med. 2019;

23. Leinonen R, Sugawara H, Shumway M, et al. The sequence read archive. Nucleic Acids Res. 2011; 39:D19-21

8 Acknowledgements

Over the last few years I've met many amazing and intelligent people who are filled with passion for science, curiosity to ask questions, and the drive to challenge themselves. I think maybe a bit of that energy and intellect might have rubbed off on me…Maybe. So thanks guys! And thanks to my family, my friends around the world, Tom, my colleagues and all who worked with me and helped me through this journey, such as you who are reading this now expecting to see your name here but won't cuz I'm too lazy to write about everyone. You're great (maybe)!

There's one person I do want to specifically mention. Without this person, this thesis wouldn't have been possible.

And, yes - that person is me.

Thank you so much past Christine for finishing your PhD despite all the temptations from reddit, Netflix, computer games and literally anything else other than work that tried to seduce you for your time. Now that this is over with, you're mentally free to complain about your next endeavor and gloat about how 'easy' the PhD was. These last years in NL were filled with fun and growth. I'm very proud and grateful for all the nice memories you've given me. Also big shout out to future Christine, the only other person who may actually read this a few years from now. Stay sweet, stay humble, stay awesome!

9 Curriculum Vitae

Birth: Dec 10 1989, Ogidi, Nigeria

Education

TU Delft, Delft, NL: Bioinformatics Delft Global Fellow	2016-2020
VU/University of Amsterdam, Amsterdam; MSc: Bioinformatics VU Fellowship Scholar	2014-2016
Duke University, Durham NC; BSc: Biology, Minor: Chemistry	2007-2011

Experience

illumina, San Diego, USA Bioinformatics Scientist	Sep 2021 - current
Amsterdam Data Collective, Amsterdam, NL Data Science Consultant	March – Aug 2021
Chinese University of Hong Kong, Hong Kong Graduate intern	Jan – July 2016
ETH Zurich, Basel CH Graduate intern	Oct – Dec 2015
Diagnostics for All, Cambridge, MA, USA Research Associate	2012-2014

10 Publications

Cao, Q., Anyansi, C., Hu, X., Xu, L., Xiong, L., Tang, W., ... & Yip, K. Y. (2017). Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. *Nature genetics*, *49*(10), 1428-1436.

Anyansi, C., Keo, A., Walker, B. J., Straub, T. J., Manson, A. L., Earl, A. M., & Abeel, T. (2020). QuantTB – a method to classify mixed Mycobacterium tuberculosis infections within whole genome sequencing data. BMC Genomics, 21(1), 80.

Salazar, A. N., Nobrega, F. L., Anyansi, C., Aparicio-Maldonado, C., Costa, A. R., Haagsma, A. C., ... & Abeel, T. (2020). An educational guide for nanopore sequencing in the classroom. *PLoS computational biology*, *16*(1), e1007314.

Anyansi, C., Straub, T. J., Manson, A. L., Earl, A. M., & Abeel, T. (2020). Computational methods for strain-level microbial detection in colony and metagenome sequencing data. *Frontiers in Microbiology*, *11*, 1925.

Anyansi, C*., Deyneka, E.*, Zade, R. S. H., van Rossum, T., Costa, A. R., Aparicio-Maldonado, C., McKenzie, R. E., Haagsma, A. C., Mulders, W., Nobrega, F. L., Meehan, C. J., Brouns, S. J. ., & Abeel, T. (2020). Reconstruction of PE/PPE genes in high-quality M. tuberculosis genome assemblies reveals lineage specific distribution patterns. (In Submission).

Anyansi, C., Cohen, K. A., Manson, A. L., Earl, A. M., & Abeel, T. (2021). Whole genome sequencing reveals extensive prevalence of mixed tuberculosis infections across the globe and its role in drug heteroresistance. (In Submission).

van Dijk, L.R., Walker, B.J., Straub, T.J., Worby, C.J., Grote, A., Schreiber, H.L., Anyansi, C., Pickering, A.J., Hultgren, S.J., Manson, A.L. and Abeel, T., (2021). StrainGE: A toolkit to track and characterize low-abundance strains in complex microbial communities. *bioRxiv*.