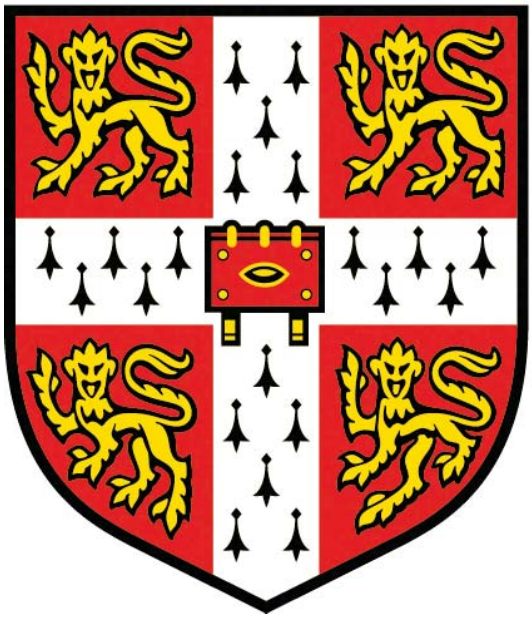
Evolutionary Genomics of Pathogenic Mycobacteria

Josephine M. Bryant



Darwin College
University of Cambridge
Wellcome Trust Sanger Institute
November 2014
This dissertation is submitted for the degree of Doctor of Philosophy

This dissertation is the result of my own work. Any work that is the outcome of work done in collaboration is specifically indicated in the text.

No part of this dissertation has been submitted for any other qualification and it does not exceed the word limit stipulated by the Biological Sciences Degree committee.

Abstract

The genus *Mycobacterium* includes many species pathogenic to human health. This thesis concentrates on two of these species: *Mycobacterium tuberculosis*, the causative agent of tuberculosis and an obligate intracellular human pathogen; and *Mycobacterium abscessus*, an environmental bacterium that can opportunistically cause respiratory and soft tissue infections in humans. Whole genome sequencing was carried out on large sample collections of these two species in order to understand how they transmit in addition to their evolutionary dynamics over small to large evolutionary scales.

For *M. tuberculosis*, a very low substitution rate of ~0.3 single nucleotide polymorphisms (SNPs) per genome per year was observed in the context of patient-patient transmission. This low genetic turnover presents challenges to our ability to use whole genome sequencing to infer direct transmission of tuberculosis, and highlights the continuing importance epidemiology will play in strengthening these inferences. Whole genome sequencing was also applied to recurrent tuberculosis disease, where patients had had a second disease episode within two years of being cured of the first. This enabled the clear differentiation of those caused by relapse and those by re-infection. In addition mixed infections were detected and deconvoluted, which would not have been possible using traditional genotyping methods. Finally the highly variable PE and PPE genome families were studied in detail using both mapping and *de novo* assembly approaches. The functions of these gene families are unknown, but they are often cell-surface associated and antigenic, so have been speculated to play a role in within-patient antigenic diversification. This analysis found that although these genes were more variable than the rest of the genome, this variability was not generated within patients, suggesting another role for these gene families.

Compared to *M. tuberculosis*, *M. abscessus* is poorly understood, with little genomic data or an understanding of population structure available prior to this study. This thesis concentrates on the infection of cystic fibrosis patients with *M. abscessus*,

which is causing concern due to its high level of antibiotic resistance and rising incidence. Whole genome sequencing was carried out on a collection from a single cystic fibrosis clinic collected over four years. For most patients, their isolates were distantly related, a pattern consistent with independent acquisition from the environment. This was expected as transmission between patients was previously assumed to be impossible or rare. Surprisingly there were some patients however who shared identical or near identical isolates, which fell into two phylogenetic clusters. This suggested transmission between patients had occurred, a conclusion supported by both epidemiological evidence and Bayesian dating methods. In addition to transmission, this dataset also provided the opportunity to capture within-patient diversity through the detection of minority variants. These minority variants were correlated with clinical outcome and treatment, revealing fluctuations in genetic diversity over time with associated changes in phenotype.

Whole genome sequencing has allowed the analysis of the evolution of two important mycobacterial pathogens over different timescales: within patient, within outbreaks and across the species. These analyses have not only provided us with greater insights into how they evolve, and at what rate, but also have had a significant clinical impact. This work has highlighted the power of the whole genome approach, especially when applied to organisms with a low mutation rate, which will be essential for furthering our understanding of mycobacteria.

Acknowledgements

This work wouldn't have been possible without Julian Parkhill, who has always made time for me, and is the most informed supervisor one could hope for in the field of bacterial genomics. Simon Harris, has not only been a good friend, but has developed many of the bioinformatic pipelines used in this thesis, and has provided advice on countless occasions. Stephen Bentley has provided guidance throughout my PhD, particularly in the early years. I'd like to thank him for encouraging me to join this group and getting me working on mycobacteria. I will miss E212 and the rest of the pathogen genomics team, who have provided a friendly and fun environment to work in. My thesis committee: James Wood, Paul Kellam and Jeff Barratt, have provided valuable advice and critique.

My work has depended on countless collaborators and support teams both within and outside Sanger. Within the Sanger, I am indebted to the DNA pipeline and Pathogen Informatics teams who have provided sequencing and bioinformatic support, allowing my research to be less about the practicalities and more about the biology. Outside Sanger, I would like to thank Ed Feil who got me interested in pathogen evolution in the first place and encouraged me to apply for this PhD; Stephen Gillespie who has always made time for me; Andres Floto and Dot Grogono who have been fantastic collaborators; and all those that provided and prepared samples, I'm aware mycobacteria can be awkward buggers to grow.

I'd like to thank my family, who has provided unwavering support and limitless pride. In particular: my dad for reminding me that science is a creative process; my mum for showing me how a toilet cistern worked at an early age, and getting me interested in the world; Liam for feeding me; and finally my grandfather Maurice for setting the bar high.

Contents

Abbreviations

CF – cystic fibrosis

CRP – C reactive protein

FEV – forced expiratory volume

HPD – higher posterior density

MDR – multi drug resistant

MGIT – mycobacterial growth indicator tube

MIC – minimum inhibitory concentration

MIRU – mycobacterial interspersed repeat unit

MLSA – multi-locus sequence analysis

MLST – multi-locus sequence typing

MTBC – *Mycobacterium tuberculosis* complex

NTM – non-tuberculous mycobacteria

PCR – polymerase chain reaction

PFGE – pulsed field gel electrophoresis

RFLP – restriction fragment length polymorphism

VNTR – Variable number tandem repeat

XDR – extensively drug resistant

1. Introduction

1.1. The genus

The bacterial genus *Mycobacterium* consists of a diverse range of both environmental and obligate intracellular bacteria. They are characterised by a distinctive waxy cell wall containing mycolic acid, distinguishing them from the rest of the Actinobacteria family. Most members are GC rich, non-motile and aerobic. Although not strictly gram positive (their thick cell wall makes them impervious to gram staining), they are classed as gram positive due to the absence of an outer cell membrane (Salyers 1995). Numerous members of the genus *Mycobacterium* are pathogenic to human health, including the causative agent of tuberculosis.

Due to the undeniable importance of tuberculosis, the taxonomy of the genus has often been skewed around it; with species being classified in terms of their relatedness to *M. tuberculosis*. This is reflected in the commonly used terms "non-tuberculous" (NTM) or "atypical" mycobacteria (Gangadharam and Jenkins 1997). Currently the genus is most often broadly divided into two: "slow growers" and "rapid growers", where the traditional division based on growth rate and molecular relationships based on 16S rRNA are in agreement (Figure 1). Rapid growers are those species that under optimal solid culture conditions grow visible colonies within seven days. The slow growers exceed this time to varying degrees. The number of valid mycobacterial species names currently stands at 169 (LPSN 2014).

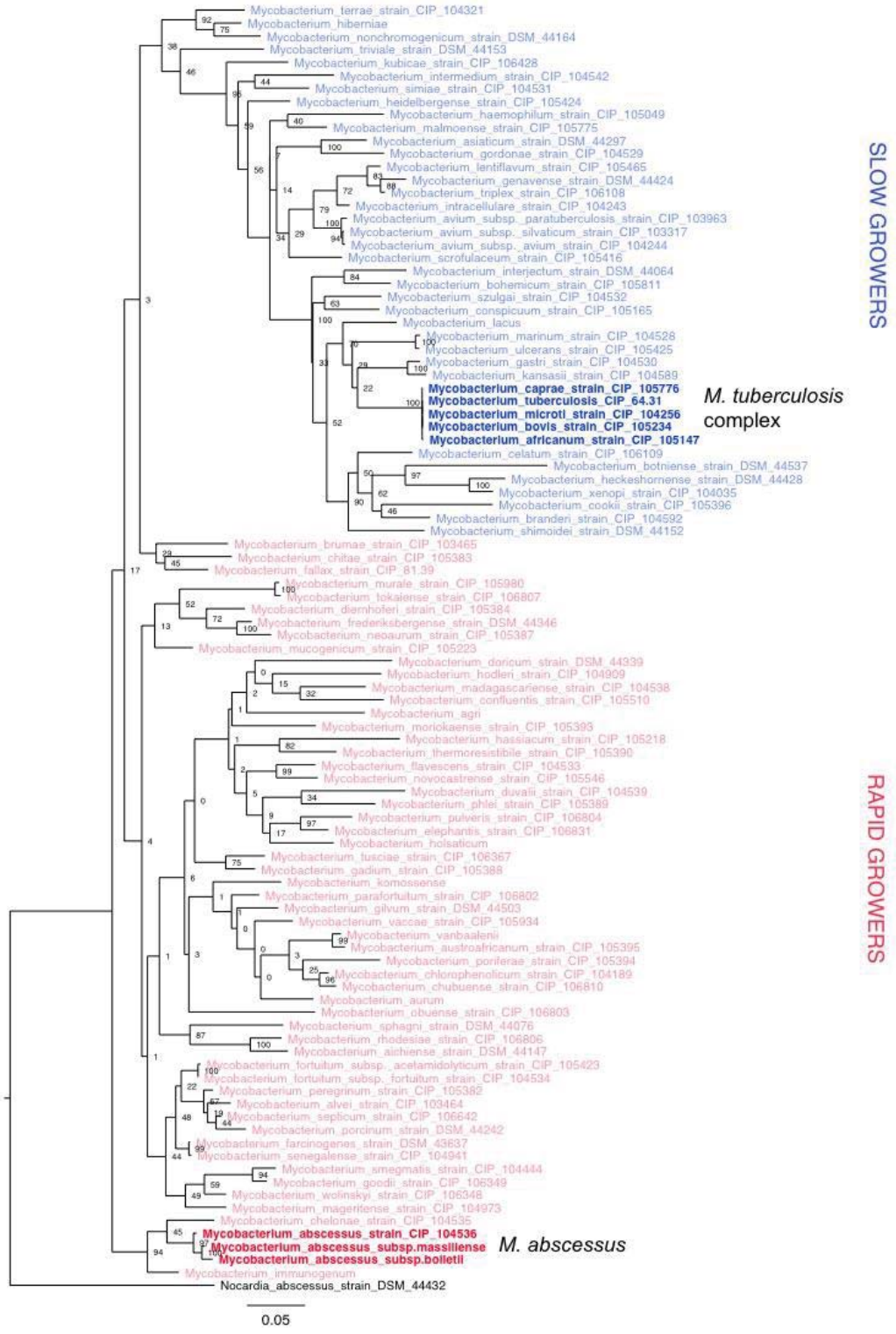


Figure 1 - Phylogeny of mycobacteria based on concatenated *Sod*, *Hsp65* and 16S sequences. Sequence accessions obtained from Devulder *et al.* (Devulder, Perouse de Montclos *et al.* 2005). Aligned with Muscle (Edgar 2004) and tree built with RAxML (Stamatakis 2006). Bootstrap support obtained from 100 trees are labeled. Rooted on out-group *Nocardia abscessus*.

The most notable members of the slow growers belong to the *M. tuberculosis* complex, which cause tuberculosis in both humans and animals. Another slow-grower is *M. ulcerans,* which is the cause of the Buruli Ulcer: a neglected tropical disease with its highest incidence in sub-Saharan Africa (World Health Organization 2013). Also of note is *M. avium subsp. paratuberculosis* which causes Johnes disease in cattle and has long been suspected (but not yet proven) to be a contributor to Crohns disease in humans (Hermon-Taylor and El-Zaatari 2004). *M. leprae* causes leprosy, a disabling disease which is still endemic in isolated pockets of the world (World Health Organization 2012). All of the known rapid growing Mycobacteria are primarily environmental, with some having the ability to become opportunistic pathogens. The most virulent and clinically relevant of these is *M. abscessus,* which can cause both wound and respiratory infections.

This dissertation will focus on two of these organisms, an obligate intracellular slow grower: *M. tuberculosis*, and the free-living rapid grower: *M. abscessus* (shown in Figure 1).

1.2. *Mycobacterium tuberculosis*

1.2.1. Pathophysiology

The life cycle of tuberculosis starts with inhalation when the infectious droplets reach the alveoli. They are quickly engulfed by the alveolar macrophages. At this point the immune system either manages to confine the mycobacteria leading to a latent asymptomatic infection, or failure can lead to an active infection. In order to control the infection, the macrophages induce production of proteolytic enzymes and cytokines that attract T lymphocytes to the site. This initial control phase can last between 2-12 weeks (Knechel 2009). If this is successful then a granuloma will eventually be formed, which is a nodular type lesion formed of T lymphocytes and macrophages intended to confine the mycobacteria. This environment is characterised by low oxygen and pH, in which the mycobacteria are able to survive in a dormant state to but is thought not to replicate. The lesion can then undergo calcification and fibrosis in order to keep the infection confined. Approximately 90% of those infected with *M. tuberculosis* are thought to maintain the infection in this dormant state for the rest of their lives (Dye and Williams 2010).

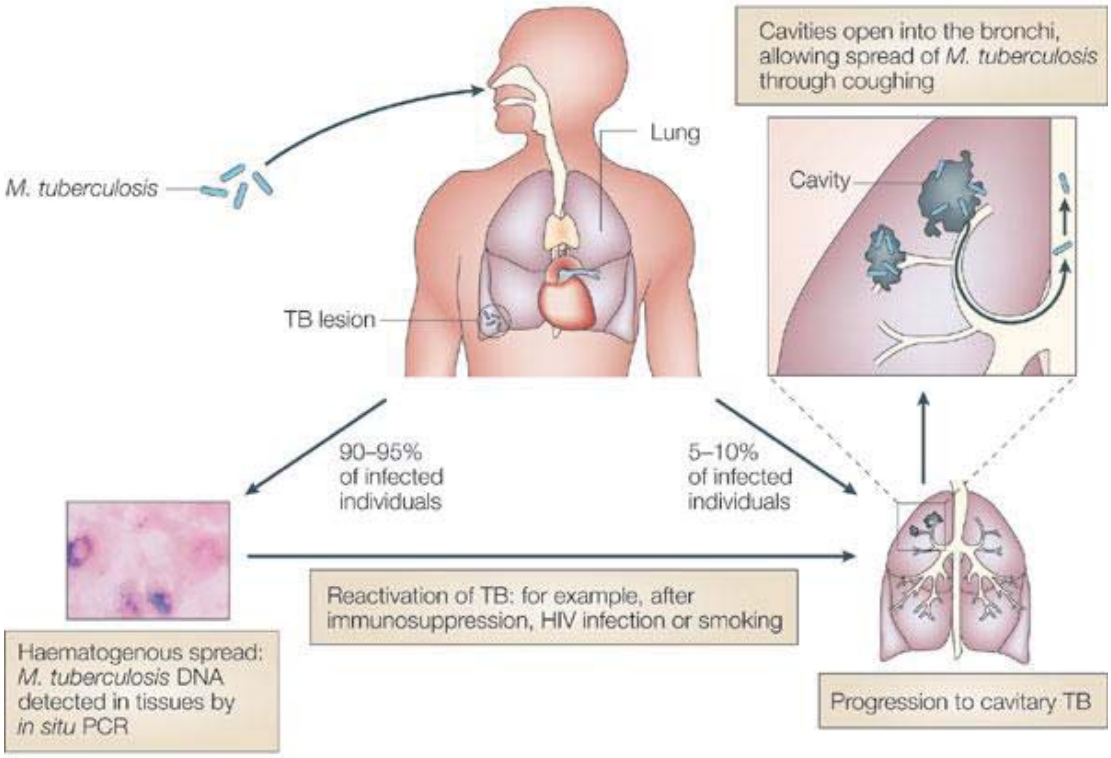


Figure 2 - Phases of human tuberculosis. After inhalation of the bacteria, there is a blood-borne stage where the immune system attempts to control the infection. In 5-10% of individuals this will lead to active or cavitary tuberculosis, which can shed *M. tuberculosis* allowing ongoing transmission through aerosol production. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Immunology, (Rook, Dheda *et al*. 2005), copyright 2005.

For the 10% that go onto develop active disease, the granuloma fails to contain the bacilli allowing them to spread to a bronchus or nearby blood vessel (Knechel 2009). This allows the infection to spread throughout the respiratory system where progressive lung damage occurs through the formation of cavities (Figure 2). In some cases it spreads to other organs such as the lymphatic system, bones and meninges. The symptoms of early and progressive active disease can be unspecific but the most common are fatigue, weight loss and a chronic cough (Knechel 2009). If untreated 50% will be expected to die of the disease. The timing of the development of active tuberculosis can vary greatly from weeks after infection to decades after, and is most often caused by a compromised immune system that is often the result of HIV, but can also be due to other medical conditions such as diabetes and malnutrition. Active disease can later become latent, and then be reactivated multiple times throughout life.

Active tuberculosis disease allows transmission to other people. This happens when droplets are coughed up from the bronchus that can remain airborne for minutes to hours allowing spread to other persons. These droplet nuclei are tiny ranging from 2–5 μm in diameter and containing as few as 1–3 cells (Riley 1957). Pioneering experiments in the 1950s on guinea pigs demonstrated that it is likely that just one infectious particle can cause an infection (Riley, Mills *et al*. 1995). In addition, more recent work on macaques demonstrated that most granulomatous lesions are established by just one bacterium (Lin, Ford *et al*. 2014). These small infectious doses demonstrate the potential ease at which this pathogen can transmit.

1.2.2. Historical perspective

Tuberculosis is considered an ancient disease, which may have co-existed with humans throughout our evolutionary history. Evidence for tuberculosis-like disease, confirmed by both morphological and molecular methods, has been found in skeletons dating to the Neolithic era, approximately 9,000 years before the present (ypb) in the Eastern Mediterranean (Hershkovitz, Donoghue *et al*. 2008). However, some estimates place the origin of the disease much earlier: 70,000 ybp when humans first started emerging from Africa (Comas, Coscolla *et al*. 2013).

Tuberculosis is thought to have killed more people than any other microbial disease throughout history (Daniel 2006). Its significant impact on human society is reflected by its many names. Consumption (or Phthisis in Greek), first described by Hippocrates, refers to the "wasting away" and weight loss experienced by sufferers (Smith 2003). During the epidemics that spread throughout Europe during the 17th and 18th centuries the term "White Plague" was used (Zumla, Mwaba *et al*. 2009), which presumably referred to the pale complexion sufferers developed. Incidence is thought to have reached its peak during the 19th century when a quarter of Europeans are thought to have died (Smith 2003). It is against this backdrop that Robert Koch made his famous presentation to the Physiological society of Berlin in 1882, where he demonstrated that the tubercle was the causative agent of tuberculosis. Not only was this one of the first pathogenic bacteria to be described but he also established

"Koch's postulates", which set the standard of infectious disease etiology, still relevant today (Daniel 2006).

With the advent of antibiotics and improved public health measures, many in the western world have considered tuberculosis a disease of the past. Incidence declined gradually during the early and mid 19th century almost until the present day (Daniel 2006), although the exact reasons for this remain unclear. Despite this, a third of the population is thought to be infected, with an estimated 1.3 million dying in 2012 (WHO 2013). In addition, a deadly combination of HIV and antibiotic resistance has raised this threat to both the developed and developing worlds. Southern Africa is particularly affected. For example in Swaziland, where 1 in 3 are HIV positive (Bicego, Nkambule *et al*. 2013), the proportion of new TB cases that are multi-drug resistant (MDR) has increased from 0.9 to 7.7% between 1995 and 2009 (WHO 2012). Extensively-drug resistant (XDR) tuberculosis (defined as multidrug-resistant disease with resistance to a fluoroquinolone plus a second-line injectable drug) is also becoming a considerable threat, resulting in extremely poor treatment outcomes. In a recent study of XDR tuberculosis in South Africa, 46% of patients had died after a 2 year follow-up (Pietersen, Ignatius *et al*. 2014); the same outcome we would expect without treatment at all. Efforts are desperately needed to prevent further resistance (leading to totally drug resistant strains (Klopper, Warren *et al*. 2013)) or to develop new drugs.

1.3. *Mycobacterium abscessus*

1.3.1. Pathophysiology

Like most NTM, *M. abscessus* is primarily an environmental bacterium, but can cause opportunistic infections in humans. These can take the form of flesh wound infections or pulmonary disease in both immuno-compromised and immuno-competent individuals (Medjahed, Gaillard *et al*. 2010). Both types of infections can be difficult to treat due to their ability to form bio-films and their natural resistance to many antibiotics, including all frontline tuberculosis drugs (Medjahed, Gaillard *et al*. 2010). It is not known where in the environment *M. abscessus* naturally resides, although the frequency of outbreaks associated with water (Dytoc, Honish et al. 2005, Nakanaga, Hoshino et al. 2011, Wertman, Miller et al. 2011) suggests that they thrive in an

aquatic environment. Further evidence for this comes from the seven cases of infection that were identified in the wounds of trauma patients after the tsunami catastrophe in Thailand in 2004 (Appelgren, Farnebo *et al*. 2008). It is likely that they replicate intracellularly within amoebae, as demonstrated experimentally (Adekambi, Reynaud-Gaubert *et al*. 2004).

Compared to *M. tuberculosis*, relatively little is known about *M. abscessus* pathogenesis and transmission, although there are some obvious similarities when considering pulmonary disease. Firstly *M. abscessus* is thought to be transmitted via inhalation of droplet nuclei (Falkinham 2003), that can be produced by natural sources such as rivers and streams (or more controversially humans). Secondly *M. abscessus* infections are granulomatous, where in a process similar to tuberculosis, the macrophages and T lymphocytes attempt to confine the bacteria to a lesion (Ordway, Henao-Tamayo *et al*. 2008). However, unlike *M. tuberculosis, M. abscessus* isn't an obligate pathogen. This may be why it has two morphotypes which it can spontaneously switch between; the rough type is thought to be more invasive and adapted for human infection whereas the smooth is considered less virulent. Once *M. abscessus* infects a person, it is thought to be at an evolutionary dead-end as NTMs are widely considered non-contagious. Prior to the work described in this thesis, there was only limited evidence for transmission between humans (Aitken, Limaye *et al*. 2012) and the medical consensus was that it was a rare or impossible occurrence (National Jewish Health. 2014).

1.3.2. Historical perspective

M. abscessus is the most pathogenic of the rapidly-growing Mycobacteria (Weiss and Glassroth 2012). It was only recognised as its own species in 1992 through DNA hybridization experiments (Kusunoki and Ezaki 1992); prior to then it was classed under the *M. chelonae* species, which has an identical ribosomal RNA 16S sequence. It is now clear that both the pathogenic (van Ingen, de Zwaan *et al*. 2009) and drug resistance potential (Nash, Brown-Elliott *et al*. 2009) of these organisms is very different.

The most significant cause for concern for *M. abscessus*, is its frequent isolation from cystic fibrosis patients, where it can cause chronic infections which can be extremely

difficult to treat. For this reason, most infections are never completely cleared through antibiotics alone and may require surgery. For example, in one patient group it was found that seven out of ten patients required surgery to clear infection (Griffith 2003). Prior to 1990 *M. abscessus (previously known as M. chelonae subsp. abscessus)* was rarely isolated from cystic fibrosis sputum, however now it is one of the top Mycobacteria causing disease in cystic fibrosis patients. For example in Paris, amongst the 9.8% patients infected with a NTM, 51.7% were due to *M. abscessus* (Sermet-Gaudelus, Le Bourgeois *et al*. 2003). Worryingly, its incidence is thought to be on the rise globally as shown by studies in Taiwan, United States, Australia and Israel (Lai, Tan et al. 2010, Prevots, Shaw et al. 2010, Thomson 2010, Bar-On, Mussaffi et al. 2014). The reason for this increase is unknown but could possibly be ascribed to a number of reasons including improved diagnostics, increased use of inhaled antibiotics (Renna, Schaffner *et al*. 2011) or transmission between patients.

1.4. Understanding the population structure of Mycobacteria

1.4.1. The *M. tuberculosis* complex

The species *M. tuberculosis* belongs to the "*M. tuberculosis* complex" (MTBC), which is a group of closely related species all with the ability to cause tuberculosis disease in animals. Although currently defined as different species, in one sense they fall short of the minimum standard to be considered true species (5% nucleotide divergence). Despite this there are clear phenotypic and epidemiological differences between the members of the complex. *M. tuberculosis* is strictly a human pathogen whereas *M. bovis* can infect a wide range of animals, but of primary concern is its burden in cattle. *M. africanum* is primarily found in humans but is unusual as it seems to be restricted geographically to West Africa. It is suspected this geographical restriction is a result of an animal population acting as a reservoir for the species (Bentley, Comas *et al*. 2012). *M. canettii* is the most divergent species, differing from *M. tuberculosis* by at least 2%. It has an unusual smooth colony morphology and a lower level of virulence in animal models (Supply, Marceau *et al*. 2013).

The completion of the first *M. tuberculosis* reference genome (Cole, Brosch *et al*. 1998) provided the opportunity to use DNA microarray technology to detect large sequence polymorphisms (LSPs). These LSPs were used as markers to reflect the

deep evolutionary relationships between members of the complex, in addition to the presence of distinct but more recent lineages within the species. Crucially they provided evidence that laid to rest the commonly proposed idea that human tuberculosis evolved from a bovine progenitor; as *M. bovis* was found to be more recently derived than human strains (Brosch 2002).

Our knowledge of the MTBC was furthered by sequence-based analyses of genes (Hershberg, Lipatov *et al*. 2008), and more recently by whole genomes (Comas, Coscolla *et al*. 2013). This revealed the presence of seven human lineages, and one animal lineage, which includes *M. bovis* (Figure 3). *M. africanum* is split into two distinct lineages, termed West African 1 and 2. The other lineages are comprised of geographically structured *M. tuberculosis* strains. Lineage 4, termed the Euro-American lineage is the most wide-spread and commonly isolated (Comas and Gagneux 2009) and Lineage 2, also known as the East-Asian lineage, is split into Beijing and non-Beijing strains. The Beijing clone is of particular concern as it is typically highly drug resistant and has recently spread from East-Asia (van Soolingen, Qian *et al*. 1995) into Eastern-Europe (Casali, Nikolayevskyy *et al*. 2012).

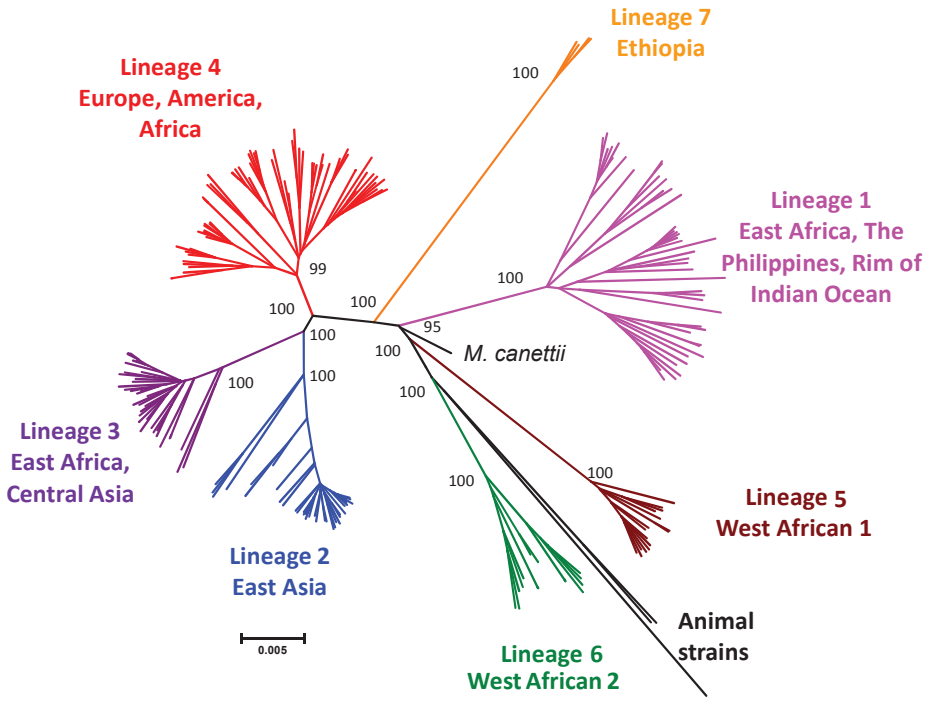


Figure 3 – Maximum likelihood phylogeny of the M. tuberculosis complex (MTBC) based on whole genome sequences. Adapted by permission from Macmillan Publishers Ltd: Nature Genetics, Comas *et. al.*, copyright 2013. Bootstrap support for the lineages are shown. The branch leading to *M. canettii* has been shortened for illustration purposes.

In addition these early studies indicated that the MTBC had a highly clonal population structure (Hershberg, Lipatov *et al*. 2008) and that there was an absence of inter-genomic recombination occurring within the complex. A recent study using whole genome sequences suggested the presence of frequent recombination within the complex (Namouchi, Didelot *et al*. 2012), however the frequency of recombination events correlates with *de novo* assembly quality (see Figure 44 Appendix 9.1) suggesting that this finding may be erroneous (manuscript in preparation). *M. canettii* is an exception where there is some evidence of recombination both within itself and with other members (Supply, Marceau *et al*. 2013). The absence of recombination in the rest of the complex is currently unexplained, but could possibly be due to a loss of the molecular mechanisms required or a lack of opportunity due to the nature of its lifestyle.

The development of molecular techniques to differentiate strains within a species has been useful as both a public health and a research tool. Over the last two decades several techniques have been developed utilizing the most variable loci in the *M. tuberculosis* genome. The first typing method developed was based on restriction fragment length polymorphism (RFLP) analysis using the insertion sequence element IS6110 as a probe (van Embden, Cave *et al*. 1993). Another commonly used technique, spoligotyping, targets specific repeat sequences found in multiple copies at a single locus in the *M. tuberculosis* genome (the direct repeat locus) using a DNA probe (Kamerbeek, Schouls *et al*. 1997). Variable number of tandem repeat (VNTR) typing is the most recently developed method, based on the presence and number of mycobacterial interspersed repeat loci (MIRU) and is currently recognised as the gold standard (Supply, Allix et al. 2006, de Beer, Kremer et al. 2012). The three methods vary in their reliability, resolution and length of time they take; but there is no clear winner with different laboratories across the world preferring either one method or employing all of them at once.

1.4.2. *Mycobacterium abscessus*

Since *M. abscessus* was given its own species designation (Kusunoki and Ezaki 1992), several different genotyping techniques have been used to probe within-species

diversity. Multi-locus sequence analysis (MLSA) or typing (MLST) are techniques based upon sequence analysis of housekeeping genes (usually seven) (Maiden, Bygraves *et al*. 1998). These are chosen on the basis that they are less likely to have recombined and more likely to be fully intact. MLSA in addition to single gene sequencing analyses have detected the presence of two or three subspecies (or species) named: *M. abscessus* subsp. *abscessus* (*M. abscessus* sensu stricto), *M. abscessus* subsp. *bolletii* (*M. bolletii*), and *M. abscessus* subsp. *massiliense* (*M. massiliense*) (Macheras, Roux et al. 2011, Macheras, Konjek et al. 2013). However, the support for these three subspecies is poor (Figure 4) with incongruence between the genes, which may be due to recombination. One study proposed that *M. a. massiliense* and *M. a. bolletii* should be combined on the basis of incomplete separation via DNA hybridization methods (Leao, Tortoli *et al*. 2011). However, there has been a lack of acceptance for this in the field, and clarity on the taxonomic status of these subspecies is only likely to be provided by whole genome comparisons.

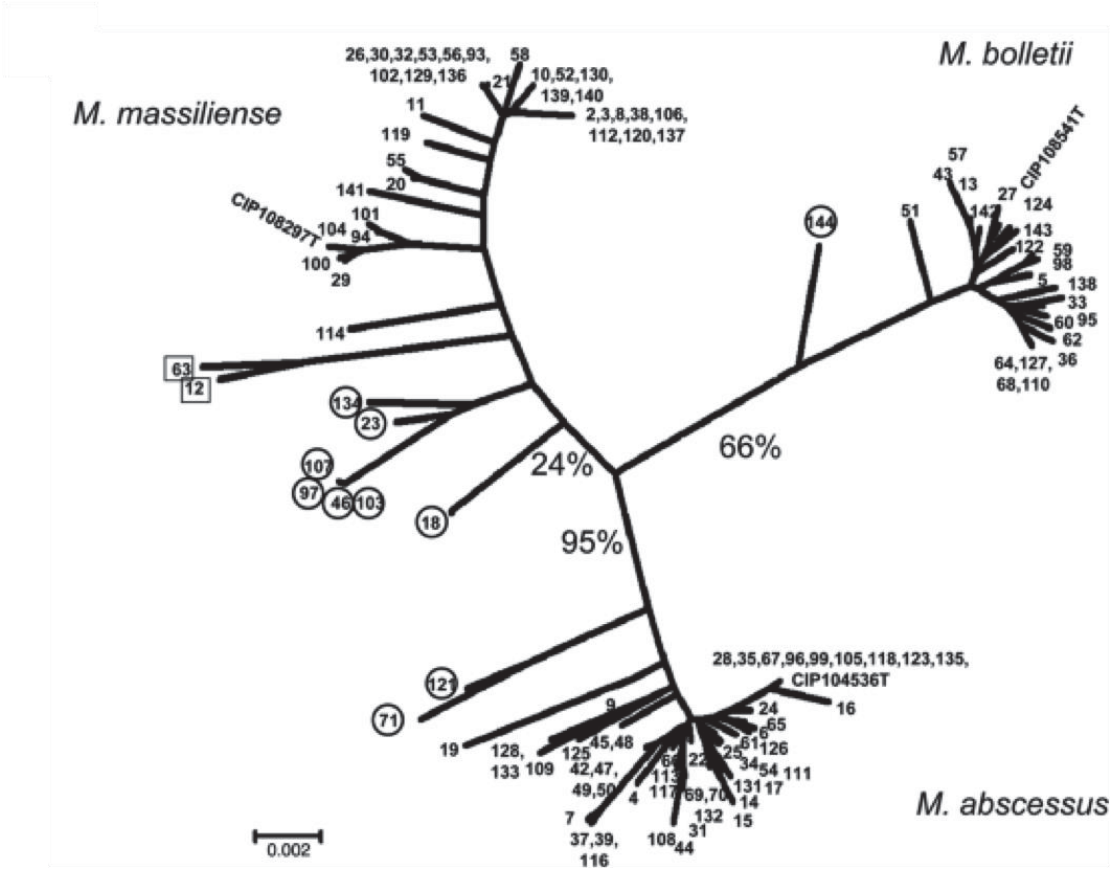


Figure 4 - MLSA based phylogeny of *M. abscessus* strains. Percentages indicate bootstrap support. Figure adapted from Macheras *et al*. (2011) with permission from the American Society for Microbiology.

Genotyping has also been used in a clinical setting to investigate the possibility of point source outbreaks (Koh, Song et al. 2010, Matsumoto, Chimara et al. 2011) or patient-patient transmission (Aitken, Limaye *et al*. 2012). In these cases methods were based on amplification of random repeat elements (Cangelosi, Freeman *et al*. 2004) or pulsed field gel electrophoresis (Zhang, Yakrus *et al*. 2004) which is based on the separation of a restriction digest of genomic DNA. These methods have been chosen as they are thought to target more variable sequences in order to provide more resolution than the sequence-based method previous described.

1.4.3. Limitations of current genotyping methods

Despite the undeniable usefulness of these genotyping techniques, it is equally undeniable that they have their limitations. By design, these loci are at the extremes of variation so are unrepresentative of the genome as a whole. So estimates of substitution rates cannot be made, and phylogenies based upon them can only provide us with a basic idea of relatedness, and are difficult to resolve with temporal information. A greater issue for public health applications, is that they can also lack resolution. This was demonstrated for RFLP typing of two *M. tuberculosis* isolates, which despite having an identical profile, differed by 130 single nucleotide polymorphisms (SNPs) detected by whole genome sequencing. What makes this particularly significant is that they had different drug resistance profiles showing that this lack of resolution can have a clinical relevance (Niemann, Koser *et al*. 2009). Likewise, isolates with identical DNA fingerprints may not always be epidemiologically linked. For example Gardy *et al.* found that for a tuberculosis outbreak in Canada, the VNTR typing data suggested it was clonal, whereas whole genome sequencing data revealed that there were in fact two concurrent outbreaks (Gardy, Johnston *et al*. 2011).

When a bacterium readily recombines its DNA, as is the case for *M. abscessus*, sequence based techniques such as MLST are further weakened. Phylogenetic trees generated from different genes can be conflicting, presumably due to inter-genomic recombination. It is only when genome-wide information is available that we can start to pick apart variants generated by recombination and those that represent the steady

vertical generation and inheritance of SNPs, as demonstrated for *Streptococcus pneumoniae* (Croucher, Harris *et al*. 2011).

1.5. Whole genome sequencing

1.5.1. Illumina Sequencing technology

Efficient methods for DNA sequencing were first developed in the 1970s through the Maxam and Gilbert method where radioactively labeled DNA was chemically degraded or through a dideoxy chain terminator method employed by Frederick Sanger. The chain terminator method was rapidly employed, and was used for most of the human genome project, automated by capillary electrophoresis (Liu, Li *et al*. 2012). Although elegant, this method is slow and time consuming. In the last decade we have witnessed an explosion in the so called "next generation" sequencing technologies which have allowed sequencing to become more high throughput and affordable. There are a plethora of technologies currently available including but not restricted to: Roche 454 the first commercially successful sequencing system; SOLiD: a high-throughput platform; IonTorrent: for small scale applications (Liu, Li *et al*. 2012), and PacBio which is designed to produce long reads. The Illumina platforms currently dominate the high-throughput sequencing market, and have been used for the majority of the sequencing carried out for this thesis, so will be discussed in greater detail.

Illumina sequencing is similar to the Sanger method in that it is based on a sequencing-by-synthesis approach, where a polymerase is used to synthesise a complementary strand to the single stranded target DNA with terminator nucleotides used to halt the synthesis. However the Illumina technology utilises reversible terminators so that the chain termination process is not permanent, and synthesis can continue after each base is detected. Fluorescently tagged nucleotides are used to determine which base is being incorporated as the synthesis proceeds one base at a time. In order to achieve this, "libraries" of the target DNA need to be prepared. First the genomic DNA is fragmented using nebulisation or sonication, with an aim to produce lots of overlapping fragments within a specific size range (for bacterial genomes this is most often 300-500bp). Adaptors are attached to the fragments which serve four functions: ligation to the flowcell, as primers for PCR amplification,

sequencing primer binding sites and as index tags to allow multiplexing of multiple libraries in a single run. After adaptor ligation a PCR step is then typically used to enrich for DNA fragments with the adaptors in the correct orientation. The DNA is then denatured to produce single strands, which are then ligated to a flowcell, where each fragment is amplified to form clusters of clonal DNA, which will increase the intensity of the fluorescent signal. The sequencing reaction is carried out with modified versions of the four nucleotides (dATP, dGTP, dCTP, dTTP) which each have a different cleavable fluorescent dye and blocking group. These allow the reaction to proceed one base at a time controlled by cleavage of the blocking group. When a base is incorporated as complementary to the template strand the fluorescent dye is photographed and then removed. This allows the sequence of the millions of DNA fragments to be determined at once; one base at a time (Liu, Li *et al*. 2012). With the current platforms (HiSeq2000, HiSeq2500) this will proceed for 100-150bp at a high quality, with error rates typically less than 0.4% (Quail, Smith et al. 2012). More information can be gleaned by carrying out paired end sequencing where sequencing is initiated from both ends of the fragment, instead of just one. This not only provides more reads but also positional information as we know roughly how far apart those reads should be in the genome based on the fragment sizes that were originally produced.

1.5.2. Bacterial Genetics to Bacterial Genomics

When the Human Genome Project was completed in 2003 (Collins, Morgan *et al*. 2003), it had taken 13 years and cost 3.8 billion dollars (Tripp and Grueber 2011). Advances in sequencing technology since then have made it possible to sequence an entire human genome in a few days, costing a few thousand dollars. As impressive as this is, bacterial genomes are megabases long as opposed to gigabases, which if run a similar fashion would produce large amounts of data, with an extremely high depth of coverage (>1000x) unrequired by most studies. Multiplexing has allowed microbiologists to utilise this technology, where unique tag sequences are added to the adaptors of sequencing libraries, meaning that the samples can be pooled into one run and then deconvoluted at the analysis stage. With the current technology (e.g. Illumina HiSeq) this means that 96 samples can be sequenced in one lane of a flow cell, making the technology not only very high throughput but also affordable.

The power of this approach was first demonstrated for a single clone of methicillin resistant *Staphylococcus aureus* (MRSA). This clone is discriminated from the rest of *S. aureus* by having MLST type 239, but within this clone MLST provides little discriminatory power. In the first study of its kind, Illumina sequencing was applied to a collection of 63 MRSA ST239 isolates representing both a broad global collection in addition to a focused dataset from one hospital in Thailand (Harris, Feil *et al*. 2010). The study was originally designed to just be a preliminary assessment of the technology, but the resultant phylogeny revealed a surprising amount of resolution. These fine-scale relationships between isolates revealed both localised transmission between wards in the hospital and also inter-continental transmission events, including a possible source of a London outbreak.

Several years on, the development of bench top sequencers has allowed these kind of retrospective studies to start becoming prospective ones. Several further studies on *Clostridium difficile* (Eyre, Golubchik *et al*. 2012) and MRSA (Eyre, Golubchik et al. 2012, Koser, Holden et al. 2012) have demonstrated this. Whole genome sequencing is starting to become a clinical reality, and the reasons for this are multi-faceted. The first reason is the high level of resolution it provides compared with the traditional genotyping technologies. This information will be valuable to infection control teams in hospitals, allowing them to make interventions; as elegantly demonstrated again for MRSA (Harris, Cartwright *et al*. 2013). It is also of great interest to academic researchers who can use this resolution to learn more about population structure, mutation rates and the evolutionary processes that drive them. Using genome wide single nucleotide polymorphisms (SNPs) as the basis for these kinds of studies means we can start to understand temporal parameters, as these units of variation are likely to be more clock-like than those studied using traditional genotyping techniques. The second major advantage of whole genome sequencing is that it can provide information on variants other than SNPs. Both mapping and *de novo* assembly approaches can be used to detect deletions, insertions and the acquisition of horizontally transferred elements such as genomic islands and mobile genetic elements (MGEs). This allows us to understand the evolutionary dynamics of these elements and how they impact on pathogenicity and antibiotic resistance. Finally whole genome sequencing is becoming cheaper, meaning that irrespective of all the

other advantages it provides, it can equally be as, if not more, economically viable than current techniques.

Whole genome sequencing does come with its challenges however. The raw data is cumbersome and difficult to analyse without the expertise and bioinformatic pipelines in place. In addition there are data storage and computing power requirements that can be a barrier to smaller laboratories. It could be considered that one of the most pressing issues however is interpretation. Once we have all this data, what does it mean and what can it tell us about transmission, antibiotic resistance and virulence for example? Each pathogen comes with its own challenges, with differing mutation rates, mobile elements and difficulties in analysis. The number of un-annotated genes we know nothing about is vast. Many studies are required in order to become confident in what the data is telling us and its strengths and limitations when applied clinically across a wide range of pathogens.

1.5.3. Application of whole genome sequencing to Mycobacteria

M. tuberculosis has attracted many large-scale sequencing projects, as reflected by the sheer number of sequences deposited in the European Nucleotide archive (currently 7,393, accessed 13/01/14). This section will summarise the major advances made in the field, and the gaps in our knowledge that still remain that could possibly be addressed using whole genome sequencing.

The first whole genome sequencing study of a *M. tuberculosis* outbreak was published in 2011 (Gardy, Johnston *et al*. 2011), which described a putative outbreak involving 32 people in British Columbia, Canada. All the isolates from the outbreak were identical using MIRU-VNTR, but could be differentiated using the genome-wide data. The resultant phylogeny was combined with detailed epidemiological contact tracing, which together suggested that what was thought to be a single outbreak was in fact two concomitant outbreaks. They also predicted the presence of a super-spreader: an individual with a particularly high ability to transmit to others. Similar studies on other tuberculosis outbreaks (Roetzer, Diel *et al*. 2013) or recent transmission within the UK area (Casali, Nikolayevskyy et al. 2012, Walker, Ip et al. 2013) and Russia (Casali, Nikolayevskyy *et al*. 2012) have provided further evidence

to support the advantages of whole genome sequencing in determining transmission over previous techniques.

Whole genome sequencing has also enabled significant progress in our understanding of antibiotic resistance. Work by Sebastian Gagneux and colleagues provided the first convincing evidence for the existence of compensatory mutations in tuberculosis (Comas, Borrell *et al*. 2011), which ameliorate the fitness cost associated with drug resistance mutations. Furthermore our knowledge of possible drug resistance causing mutations have also been expanded either through the identification of convergent mutations (Farhat, Shapiro *et al*. 2013), or through more complex methods that identify evidence of diversifying selection in both genic and intergenic regions associated with drug resistant strains (Zhang, Li *et al*. 2013). Of particular note is a study utilising 1,000 strains from a single time-point and area in Russia where MDR tuberculosis is a particular problem. They found that the same drug resistance mutations had evolved independently several times, and that particularly dominant clades had gained putative compensatory mutations, which were likely to explain their success (Casali, Nikolayevskyy *et al*. 2014). Although these kinds of in-depth large-scale studies are likely to achieve greater progress in our understanding of antibiotic resistance, it is currently unknown how much we have explained, and how much it is possible to explain. All acquired resistance in *M. tuberculosis* is thought to be via *de novo* chromosomal mutation, so in principle whole genome sequencing should be able to make significant advances. However, the underlying processes contributing to resistance are likely to be complex and multifactorial, as there is increasing evidence for the step-wise accumulation of mutations on the path to resistance (Safi, Lingaraju *et al*. 2013), and for hetero-resistance (Rinder 2001, Sun, Luo et al. 2012). It is particularly pertinent that future studies are carried out in Africa, where we have very little understanding of the basis and transmission of resistance, despite it making up a large proportion of the global burden.

In order to fully understand the substitution or mutation rate of a bacterium, it needs to be characterised over a number of different time scales – as it is known to be heavily dependent on evolutionary scale (Ho, Shapiro *et al*. 2007), with an apparently slower substitution rate observed between more distantly related bacteria (Ochman, Elwyn *et al*. 1999). So far this has been determined for several tuberculosis outbreaks

or transmission chains, where it appears to be extremely low at 0.3-0.5 SNPs per genome per year (Roetzer, Diel et al. 2013, Walker, Ip et al. 2013). Using the macaque monkey as an experimental model, a rate of 0.38 was derived, suggesting that the within-host substitution rate is highly similar (Ford, Lin *et al*. 2011). However this study involved a small number of isolates and infections, and it is unknown what the effect of antibiotic pressure and/or HIV co-infection will have on *M. tuberculosis* mutation in humans. It is also unknown whether the different *M. tuberculosis* lineages have different rates of mutation, but it has long been speculated that the Beijing lineage may have a higher mutation rate (Mestre, Luo *et al*. 2011), and that this could possibly explain the high level of antibiotic resistance in this clade. Work published last year (Ford, Shah *et al*. 2013) based on *in vitro* experiments, suggests this may be the case, but requires further investigation as it is counter to previous findings (Werngren and Hoffner 2003). Understanding the mutation rate over these different time scales may have important implications for not only antibiotic resistance, but also for our understanding of how *M. tuberculosis* evolved. It has been suggested that *M. tuberculosis* may have co-evolved with humans since they emerged out of Africa over 50,000 ybp, and the most compelling evidence for this are the similarities between human mitochondrial and *M. tuberculosis* phylogenetic trees (Comas, Coscolla *et al*. 2013). If this were the case the mutation rate of *M. tuberculosis* would need to average 0.01 SNPs per genome per year over most of its evolutionary history (Comas, Coscolla *et al*. 2013). More recently, analysis of ancient DNA from 1,000 year old Mummys from Peru dates the most-recent common ancestor of modern tuberculosis to 5,000 ybp, with a substitution rate of 0.3 SNPs per genome per year: a rate much more consistent with modern estimates (Bos and Krause in press). The age of *M. tuberculosis* is still a contentious issue, but is sure to be clarified further as whole genome sequencing is applied to a greater number of ancient DNA samples.

In nearly all of studies mentioned above, 10% of the coding genome of *M. tuberculosis* was discarded as it encodes genes belonging to the PE/PPE gene families. This is because they present difficulties to our ability to both sequence and to analyse them, due to their high GC content (up to ~85%) and their repetitive nature. These regions are of high interest due to their possible involvement in virulence (Sani, Houben *et al*. 2010), and the high level of diversity observed between isolates (Talarico, Cave et al. 2005, Talarico, Zhang et al. 2008, McEvoy, Cloete et al. 2012).

Finding a way to capture their diversity and quantify it in a high-throughput manner would enable us to understand more about this gene family and its impact on pathogenicity.

The gaps in our knowledge for *M. abscessus* are even larger. The first full genome sequence of *M. abscessus* was published in 2009 (Ripoll, Pasek *et al*. 2009), which revealed a number of interesting genes implicated in pathogenicity that may have been horizontally acquired from other cystic fibrosis pathogens. However no large scale study on the population structure or genomic diversity of this species had been investigated prior to the work described in this dissertation.

1.6. Thesis aims

The general aim of this thesis is to investigate the evolution and population genomics of two Mycobacterial species using whole genome sequencing. Specifically:

1) Understand the genome wide substitution rate of *M. tuberculosis* in the context of transmission and recurrent infection.
2) Understand the diversity of the PE and PPE genes in *M. tuberculosis* on different evolutionary scales.
3) Investigate the population structure of *M. abscessus* within a single cystic fibrosis clinic.
4) Investigate the long-term within-patient evolution of *M. abscessus* in cystic fibrosis patients.

2. Genomic diversity of *Mycobacterium tuberculosis* over short time scales

The majority of this work has been published in:

J. M. Bryant, A. C. Schurch, H. van Deutekom, S. R. Harris, J. L. de Beer, V. de Jager, K. Kremer, S. A. van Hijum, R. J. Siezen, M. Borgdorff, S. D. Bentley, J. Parkhill and D. van Soolingen (2013). "Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data." BMC infectious diseases **13**(1): 110.

Statement of contribution:

I carried out all bioinformatic analyses and interpretation. The study was initiated by JP, KK and DVS. Advice and bioinformatic training was provided by SRH. Help with interpretation and metadata was provided by ACS. Strain selection and sample preparation were carried out by the Tuberculosis Reference laboratory for the Netherlands.

2.2. Introduction

Many consider the routine clinical application of whole genome sequencing to *M. tuberculosis* to be inevitable in the near future (Koser, Ellington et al. 2012, Walker, Ip et al. 2013). Contact tracing and outbreak investigation will particularly benefit, as these currently rely on the information gathered from epidemiological investigation and molecular typing techniques to determine all possible contacts and routes of transmission for each patient. The higher level of resolution that whole genome sequencing offers could enable a more informed ability to include or refute possible links. However, in order to achieve this an understanding of the molecular clock over short time scales in the context of direct patient-patient transmission is essential. By knowing how much variation is generated within and between patients we can begin to judge whether direct transmission is likely to have occurred between individuals, given what we know about their contact.

Estimations of mutation or substitution rate are very much dependent on the evolutionary scale on which they are being measured on. The term mutation rate commonly refers to the basal *de novo* generation of mutations in the absence of selection. Instead here, where the variants detected in the context of patient-patient transmission are being sampled, the term substitution rate is used. In this context mutations have risen to a high enough frequency to be sampled (so highly deleterious mutations have been removed), but may not become permanently fixed in the *M. tuberculosis* population. Currently our knowledge of the intra-patient mutation rate (closer to basal mutation rate than inter-patient comparisons but not selectively neutral) of *M. tuberculosis* is limited to a single study based on infection experiments of macaques. Based on 15 isolates collected from four infections, an estimate of 0.39 (0.16-0.80 95% CI) single nucleotide polymorphisms (SNPs) per genome per year was estimated (Ford, Lin *et al*. 2011). More recently, an estimate of 0.5 SNPs/genome/year was made on the basis of within and between patient sampling of 93 patients from a larger study in the UK (Walker, Ip *et al*. 2012). This suggests that the intra-patient mutation rate may be similar to the inter-patient substitution rate.

This study aimed to further characterize the rate of change of *M. tuberculosis* observed between patients over short time scales, and to explore the strength of this technique to refute and confirm direct transmission links.

2.3. Methods

199 *M. tuberculosis* isolates were chosen by the Municipal Health Service in Amsterdam, comprising of isolates from 151 patients with 97 known epidemiological links between them, and another 48 patient isolates from the same RFLP clusters but with no evident epidemiological link. Isolates were chosen to represent a broad phylogenetic range, belonging to 42 RFLP clusters. All genotyping including IS*6110* RFLP, Spoligotyping and 24-locus variable number of tandem repeats (VNTR) typing were performed using standardized methods (van Embden, Cave et al. 1993, Kamerbeek, Schouls et al. 1997, Supply, Allix et al. 2006) by the Tuberculosis Reference laboratory the RIVM, the Netherlands.

The genomic DNA libraries were subjected to paired-end sequencing on the Illumina Genome Analyzer GAIIx platform. Thirty-three of the isolates were sequenced with a read length of 76 bases and the remaining 166 with a read length of 108 bases.

The raw sequencing reads were mapped to a corrected version of the H37Rv reference and variants were called as described in Methods 8.2 and 8.3. Mapping and SNP calling were also carried out independently at the Center of Molecular and Biomolecular Informatics (CMBI), Radboud University, using RoVar (Robust Variant detection in genome sequences using Next Generation Data from various platforms: Jager, B.A.M. Renckens, R.J. Siezen, and S.A.F.T. van Hijum, unpublished). The mapping results were compared using the epidemiological linked pairs as a test set. Most SNPs were found to agree except those found in regions flanking insertions. As short insertions and deletions are difficult to call in general, only SNPs were considered for all subsequent analysis. The genetic distance was calculated between epidemiologically linked pairs by comparing the SNPs called in each isolate. A SNP difference was only counted where there was high confidence in the base call in both isolates.

A maximum likelihood tree was constructed as described in Methods 8.6. Path-O-Gen was used to plot root to tip distances against time (Rambaut 2007) (see Methods 8.7). This program uses linear regression to root trees with date information at the position that is most compatible with the assumption of the presence of a molecular clock.

2.4. Results

2.4.1. Overview of mapping results

For the 199 samples, sequencing reads covered an average of 95.6% of the genome to a depth of approximately 100 fold. With respect to H37Rv, 11,879 positions had a SNP called in at least one of the isolates. A maximum-likelihood phylogeny based on the variants called, revealed four of the globally dominant lineages (Figure 5). The 97 linked pairs had a mean SNP difference of 3.42 (range of 0-149) and 37 of the pairs had no detectable SNP difference.

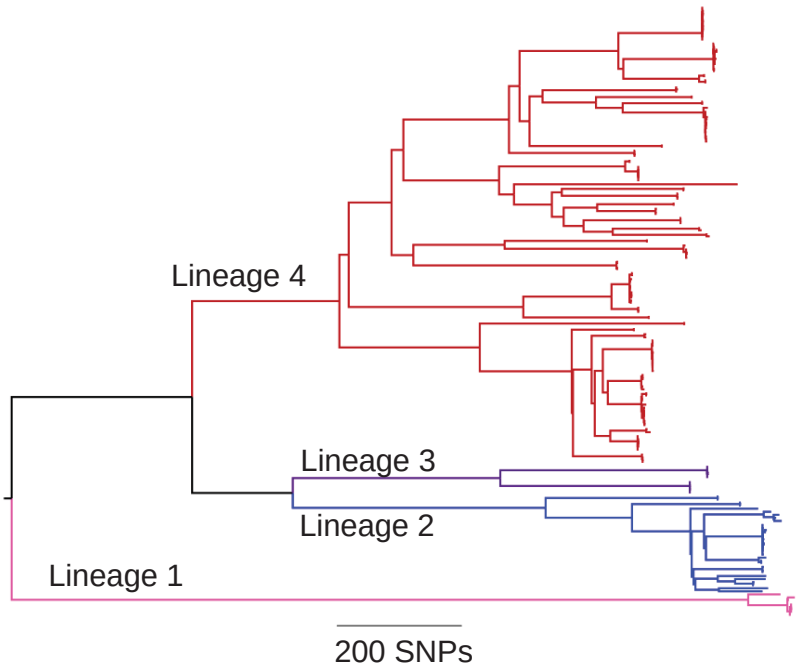


Figure 5 - Whole genome maximum likelihood phylogeny of 199 samples in dataset.

2.4.2. Homoplasic SNPs

In total, 16 homoplasic non-synonymous SNPs were identified (Table 1). Only two synonymous homoplasies were identified (Rv0161, which encodes an oxidoreductase and Rv2005c and stress related protein), suggesting that the high number of non-

synonymous homoplasies observed was unlikely to have occurred by chance alone. Five of these found in genes *rpoB* (Ramaswamy and Musser 1998), *gyr*A (Takiff, Salazar *et al*. 1994), *rrs* (Maus, Plikaytis *et al*. 2005), *kat*G (Heym, Alzari *et al*. 1995) and *emb*B (Sreevatsan, Stockbauer *et al*. 1997) have previously been associated with drug resistance. Two of the other homoplasic SNPs occurred in genes thought to be involved in pathogenicity: *ino1* (Movahedzadeh, Smith *et al*. 2004) and *opcA* (Jiang, Zhang *et al*. 2006) and three (*opcA*, Rv2082 and Rv3077) were observed in a recent study focusing on convergent evolution in *M. tuberculosis* (Farhat, Shapiro *et al*. 2013). It is likely that homoplasy has occurred in these genes due to recurring selective pressures for traits such as antibiotic resistance. This suggests that the 11 homoplasic SNPs with no ascribed function deserve further investigation, as they may possibly represent previously un-described pathogenicity, antibiotic-resistance or associated compensatory mutations.

Table 1: Homoplasic SNPs identified in this study

Gene	Number of branches	Amino acid change
gyrA	2	D94G
rpoB	2	S450L
Rrs	2	C517T
katG	5	S315T
embB	2	M306V
ino1	2	G190R
Rv0750	2	L27V
ribG	2	K30N
opcA	2	A103T
Rv1760	2	M397T
lldD2	5	V253M
Rv2082	2	L53R
Rv2709	2	E42K
Rv3077	2	R452H
Rv3463	2	G94D
aspB	2	R358Q

2.4.3. Deriving a molecular clock

For the 97 epidemiologically linked pairs, the relationship between time and the number of SNPs accumulated was investigated. Only SNPs accumulated in the secondary case in each of the linked pairs were used and SNPs found only in the

primary case isolate were excluded, as these are likely to represent either variation in the source host population that is not present in the transmitted population, or SNPs generated via laboratory passage. SNPs conferring drug resistance were also excluded (n=7), as these are likely to be subjected to a strong selection pressure and be less clock-like in the rate in which they are appear. In addition three pairs were excluded based on the phylogenetic evidence discussed below in section 2.4.4.

There was a poor correlation between the number of SNPs accumulated and the time elapsed for each patient pair (Figure 6). However, when drug resistant and sensitive pairs were plotted separately, there was an improved correlation for the sensitive pairs, but not the drug resistant pairs (Figure 6). The reason for this is unclear but a possible explanation could be based on the differing selection pressures and effective population sizes of the two groups.

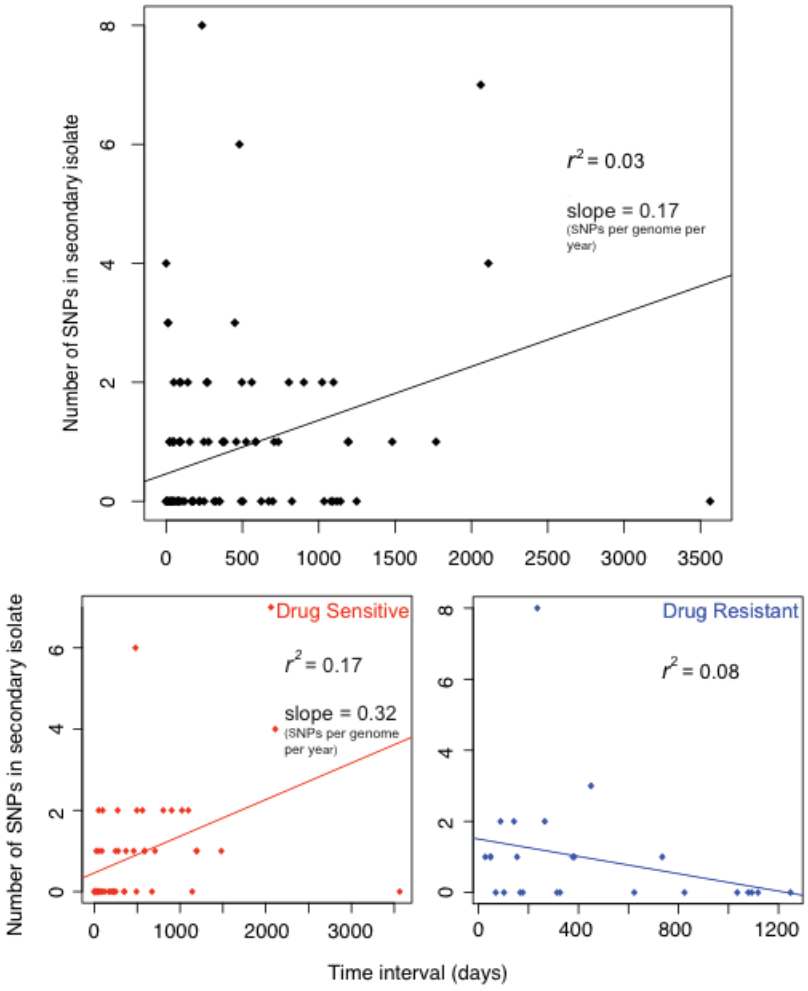


Figure 6 - Poor correlation between time and number of SNPs accumulated in the secondary case isolate for drug resistant and sensitive isolates. Correlation coefficient for linear regression models are shown

The slope of the graph provides an estimate of substitution rate, and for the sensitive isolates this was 0.32 SNPs per genome per year but with a large degree of variation around the mean reflected by an r^2 value of 0.17. This variability could reflect a number of sources of error that have to be taken into account. The first possible source of error is that the epidemiological inference could be incorrect and that direct transmission did not occur between the pairs. Secondly, there is an unknown degree of error regarding how well the date of transmission is represented by the date of isolation. Transmission is likely to have occurred prior to the isolation date (as transmission is generally considered unlikely for a patient receiving treatment), but due to the slow progression of the disease the degree of error could be very large. Finally, for 26 pairs SNPs were detected in the primary case (averaging 0.64 SNPs per pair) that were not found in the secondary, suggesting that the sampled isolate is unlikely to represent the transmitted population. This suggests that genetic diversity in the infecting population of *M. tuberculosis* exists in patients, of which only a small proportion has been sampled.

To control for the sources of error described above, a substitution rate was inferred from the entire dataset, thus not requiring assumptions about the routes of transmission. The presence of a clock-like signal was investigated using Path-O-Gen (Rambaut 2007) for the 197 samples for which a date of isolation was available. Lineage specific phenotypes have been frequently proposed (Brown, Nikolayevskyy et al. 2010, Krishnan, Malaga et al. 2011), and due to the possibility that the different lineages may have different mutation rates, the analysis was carried out per lineage. There was a complete lack of correlation between the accumulation of SNPs and time for all the lineages (Figure 7).

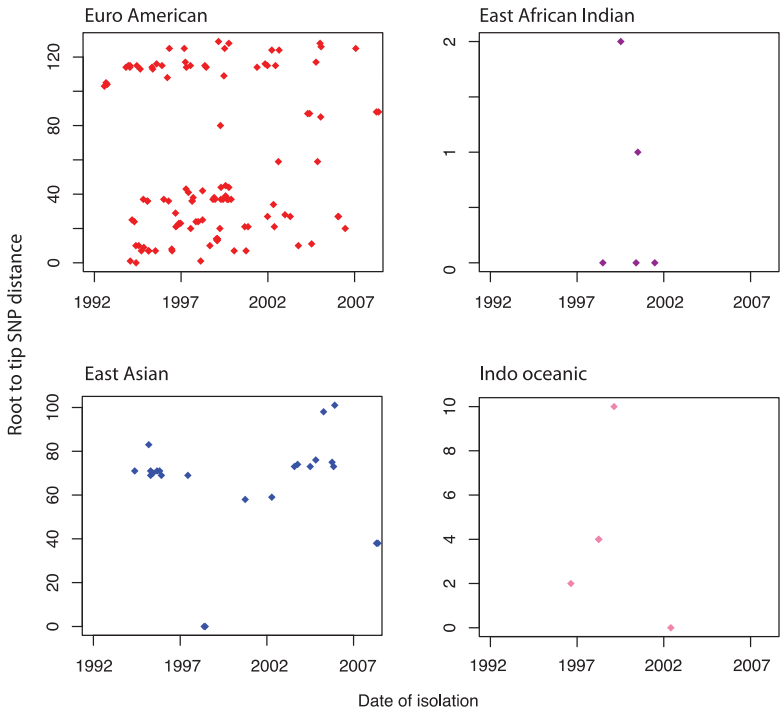


Figure 7 - Per lineage root to tip distance vs. time. The correlation coefficient of the linear regression models was poor for all lineages, with r^2 values of 0.002,0.03,0.006 and 0.06, clockwise.

This is perhaps unsurprising when the age of these lineages is considered, which would exceed thousands of years if the out-of Africa model of *M. tuberculosis* expansion is correct (Comas, Coscolla *et al*. 2013). The time dependency of the molecular clock is a well described phenomenon (Ho, Shapiro *et al*. 2007), and only the more recent variation nearer the tips of the tree is likely to be neutral and clock-like. For this reason, this analysis was then carried out on five of the largest within-lineage clusters that are more likely to represent a neutral accumulation of SNPs (Figure 8a). The linear regression slope ranged from 0.08 to 0.43 SNPs per genome per year, with this variation probably reflecting the small number of isolates and SNPs observed. When the cluster data was combined, a mean rate of 0.27 SNPs per genome per year (95% CI 0.13, 0.41) was estimated (Figure 8b). Additionally, when the age of the clusters was plotted against the number of SNPs accumulated (Figure 9), controlling for the number of isolates, a similar rate of 0.34 SNPs per genome per year was obtained.

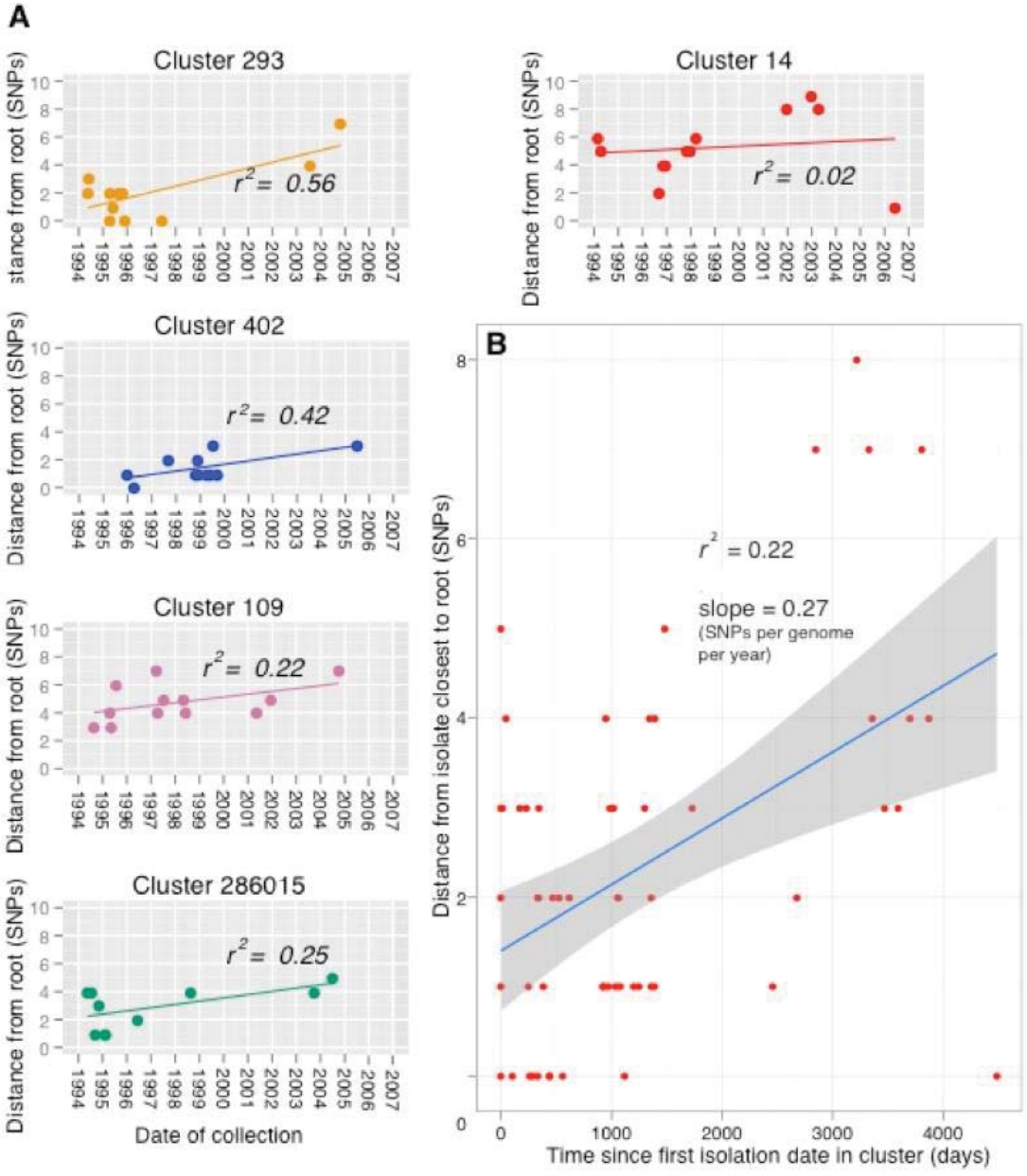


Figure 8 - Date of collection vs. root to tip SNP distance of the 5 largest clusters. B. Data combined from A. Time represents days since first isolation in the cluster. Shaded area indicates 95% confidence of linear regression model.

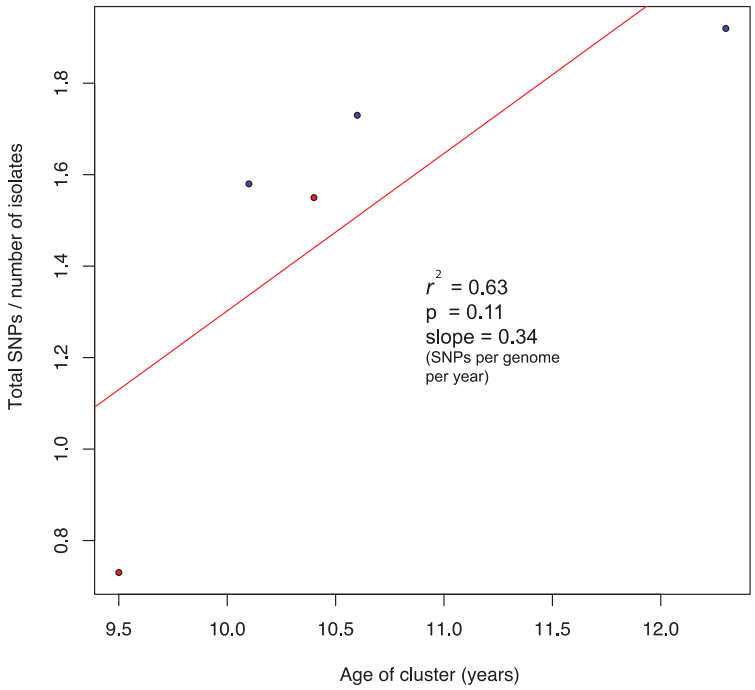


Figure 9 - Diversity vs. age for the five largest clusters. The total number of SNPs was corrected for no. of isolates in the cluster. Red indicates drug resistant clades and blue are drug sensitive.

In summary, three methods agreed on an average rate of ~0.3 SNPs per genome per year which is remarkably similar to that estimated by Ford *et al* (Ford, Lin *et al*. 2011) using the tuberculosis macaque infection model. However, the correlation co-efficient was consistently poor (maximum r^2 value of 0.6 in one case) and the level of variation observed at the isolate level was high. In addition the statistical significance of these correlations could not be assessed, as the root-to-tip distances are not independent. A more rigorous way to estimate the substitution rate is to use Bayesian coalescent analysis. However none of the datasets discussed here successfully converged using BEAST v1.7.5 (Drummond and Rambaut 2007) (see section 8.8). This is likely due to small sample size, and sampling frame and the very low rate and stochastic nature of SNP accumulation, and indicates that this estimate needs to be used with caution.

2.4.4. Using phylogeny to exclude direct transmission

Instead of looking at possible transmission events in isolation, deep sampling of a phylogenetic cluster can provide context, which can be used to make more confident inferences. Thus the structure of a phylogeny can be used to assess whether a direct transmission event is likely to have occurred. Isolates that represent a recent

transmission event are expected to be adjacent on the tree and share a most recent common ancestor, as shown in Figure 10d. If other isolates occupy the common nodes between the linked isolates in question, then this is evidence against direct transmission. This scenario was identified for two of the pairs in the study (Figure 10a and b). However, it is possible that the source case could have been carrying an infection with a diverse *M. tuberculosis* population, comprised of several sub-lineages as observed previously (Sun, Luo *et al*. 2012). In such a scenario, the entire cluster may in fact represent within patient diversity and each patient isolate is effectively a sample of this. As liquid cultures (i.e. not colony purified) were used in this study, this heterogeneity may be preserved at the variable positions. However, no evidence for this was found, and in the absence of multiple samples from each patient, this strongly suggests that these pairs do not represent direct transmission events.

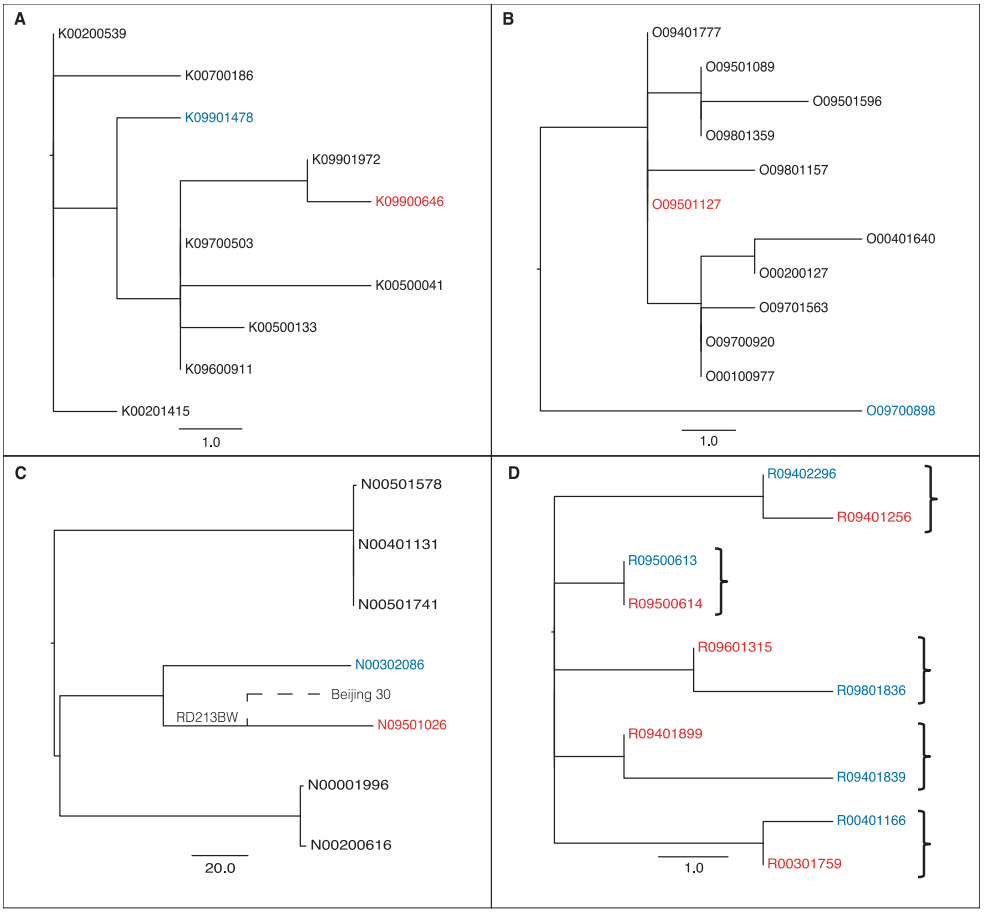


Figure 10 - Exclusion of epidemiologically linked pairs based on phylogenetic position. Red indicates primary case isolate, blue is the secondary case isolate. Maximum likelihood trees were rooted using the nearest non-clustered isolate as an outgroup. A: excluded pair 1. B: excluded pair 2. C: Excluded pair with SNP difference of 149. D: Example of expected phylogenetic positioning of direct transmission pairs, brackets indicate paired isolates.

One pair from the East Asian linage had a particularly large SNP difference of 149 (Figure 10c). The suspected source case patient lived in the same street as the suspected secondary case patient. It is unclear, however, if they were in direct contact with each other. Both isolates shared an IS*6110* RLFP pattern but their 24-locus VNTR pattern differed in 6 loci. With no detectable evidence of recombination or SNPs in possible hypermutator genes, the sequencing data was examined more closely and two independent deletions were identified: each unique to either isolate (Figure 11). The large deletion of part of the *pks1* gene found in the source isolate was found in another East Asian strain, Beijing 30, in a previous study (Tsolaki, Gagneux *et al*. 2005), suggesting that a more recent common ancestor exists than between these two isolates. This evidence along with the large SNP difference means that the possibility of recent direct transmission can be confidently excluded. In the absence of whole genome sequencing, the clear genetic separation of these isolates would have been un-detectable.

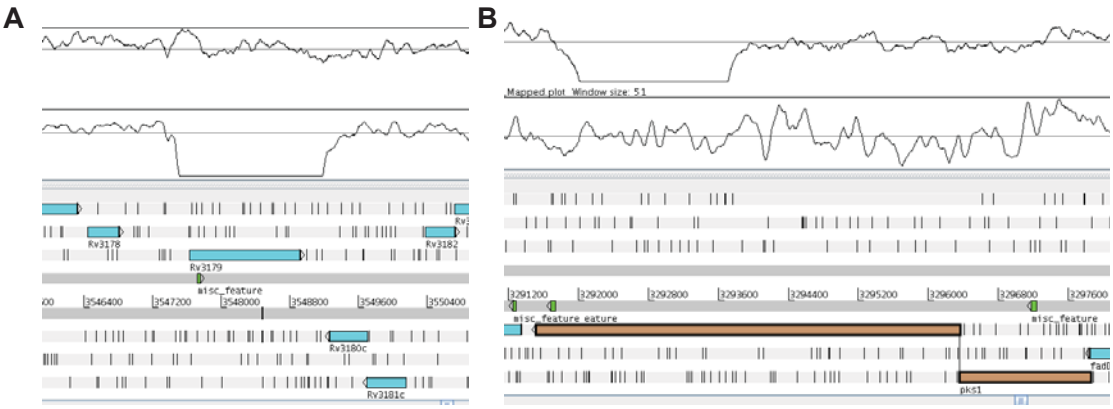


Figure 11 - Deletions identified in isolates N09501026 (top row) and N00302086 (bottom row). Plots represent mapping coverage to H37Rv reference genome. Figures adapted from Artemis (Carver, Harris *et al*. 2011). A) Deletion in Rv3179 in N00302086. B) Deletion in *pks1* in N09501026.

2.4.5. Identifying novel transmission events

In low incidence countries, identical RFLP types are often used as an indicator of possible transmission. In this dataset, 572 pairs of isolates had identical RFLP types, which had SNP distances ranging from 0-149, with a median of two SNPs. Figure 12b further confirms that the linked pair with a SNP distance of 149 is a clear outlier showing that it is distinct from the rest of the same-RFLP and epidemiologically

linked pairs. 95% of same-RFLP pairs have SNP distances under 11, indicating that in general RFLP type is a good indicator of phylogenetic relatedness. However, whole genome sequencing provides a higher resolution. For example in Figure 10d, all of the isolates in this cluster would be indistinguishable via RFLP, but at the whole genome level individual transmission events can be inferred. Figure 12b demonstrates that many pairs of isolates of the same RFLP type, with currently no known epidemiological link, have SNP distances that overlap with the range observed between the 94 linked pairs.

Strikingly, 82 pairs of these non-linked isolates, of the same RFLP type, had a SNP difference of zero. This suggests that amongst these pairs there may be previously un-detected transmission events. The range of date intervals between these pairs ranged from zero days to almost 5 years. In the absence of epidemiological evidence, and the low and variable mutation rate observed, it would be difficult to assess whether direct transmission has occurred in these cases, however this information would provide valuable evidence in a clinical setting, informing further investigation and contact tracing.

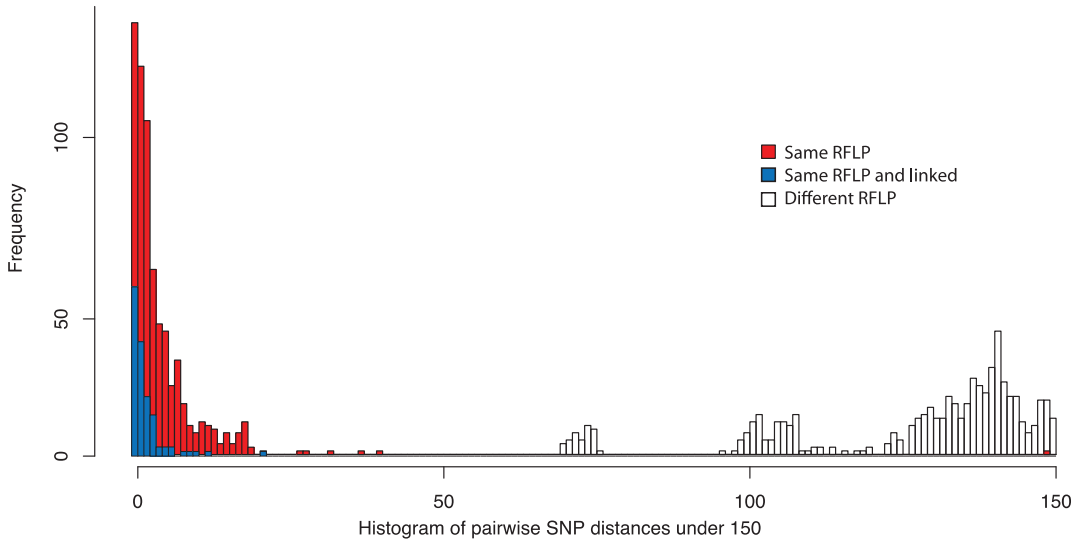


Figure 12 - Pairwise SNP differences between isolates. Only pairwise SNP distances under 150 are shown. There are many unlinked pairs that have SNP distances which overlap with the distribution of SNP distances for linked pairs.

2.5. Discussion

At the cluster level an average substitution rate of 0.3 SNPs per genome per year was estimated, which is remarkably close to estimates made in the macaque model (Ford,

Lin *et al*. 2011). This is also highly similar to the rate estimated in Oxfordshire of 0.5 (Walker, Ip *et al*. 2013), and subsequent work in Germany which concluded an average of 0.4 SNPs per genome per year (Roetzer, Diel *et al*. 2013) . This confirms the extremely low rate of accumulation of variation that characterizes *M. tuberculosis*, which is approximately 3 times and 44 times slower than that observed in *Escherichia coli* and *Staphylococcus aureus*, respectively (Didelot, Bowden *et al*. 2012).

This analysis was unable to detect a clock-like signal at a larger phylogenetic scale (at the lineage level), probably reflecting the different processes of fixation and substitution having variable influences on different parts of the evolutionary history. This confirms that the inter-patient substitution rate estimated here cannot be extended to deeper evolutionary histories, where the species-wide substitution rate is likely to be different. However, even at the intra-cluster level and between the epidemiologically linked pairs where we were able to estimate a rate, there was a large level of variation around the mean, which is in contrast to observations of other bacteria such as *Staphylococcus aureus* and *Vibrio cholerae* (Harris, Feil et al. 2010, Mutreja, Kim et al. 2011). There are a variety of factors that may have contributed to this noisy signal. Latency is common in tuberculosis infection and could result in considerable discrepancies in the apparent rate of substitution over time. However, work by Ford *et al*. showed that the substitution rate during latency in macaques was similar to that during active infection (Ford, Lin *et al*. 2011). Within host selection for factors such as drug resistance may also result in variation in the accumulation of SNPs over time. However, perhaps the most important factor is the low substitution rate itself, meaning over short time scales only a weak signal of a molecular clock can be detected.

This lack of a strong signal means that although a molecular clock may be detectable over longer time frames, it is only an aggregate measure and should be used with extreme caution when applying it to infer local transmission or date recent evolutionary events. Furthermore, while substitution rate can be used to strengthen or exclude epidemiological links, it cannot be used alone to infer direct transmission, particularly due its slow rate. There was no evidence of hyper-mutation in this dataset, and this has not been reported in clinical *M. tuberculosis* isolates to my knowledge.

However it is possible that treatment may impose selection pressures on isolates that could affect the observed rate of fixation, and this should be considered.

The Oxfordshire study concluded that a cutoff of ≤5 SNPs for cases less than three years apart may be appropriate for concluding transmission (Walker, Ip *et al*. 2013). This more in-depth analysis of the molecular clock reveals a lack of a strong signal with a high degree of variation around the mean rate, suggesting that using a simple cut-off may not be entirely appropriate for confirming transmission, but that the phylogenetic context provided by deep sampling of clusters may be more informative. This is highlighted by the fact that direct transmission between one epidemiologically linked pair separated by 5 SNPs was excluded, based on the presence of intersecting unrelated strains in the phylogeny (Figure 10a). Phylogeny and the context from other strains are important tools that can be used to further inform us on the likelihood direct transmission has taken place.

In summary, the slow molecular clock of of *M. tuberculosis* means that even at the highest resolution provided by whole genome sequencing it is still difficult to confidently affirm the inferences of transmission made by traditional epidemiological techniques. This means it is very difficult to determine transmission inclusively. However, whole genome sequencing does in some cases allow us to exclude direct transmission, by using the phylogenetic context provided by other strains. Understanding the limitations and strengths of this approach will be important for future clinical applications, and has also informed on the rest of work discussed in this dissertation.

3. Disentangling recurrent and mixed *Mycobacterium tuberculosis* infections

Based on work published in:

J. M. Bryant, S. R. Harris, J. Parkhill, R. Dawson, A. H. Diacon, P. van Helden, A. Pym, A. A. Mahayiddin, C. Chuchottaworn, I. M. Sanne, C. Louw, M. J. Boeree, M. Hoelscher, T. D. McHugh, A. L. C. Bateson, R. D. Hunt, S. Mwaigwisya, L. Wright, S. H. Gillespie and S. D. Bentley (2013). "Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study." The Lancet Respiratory Medicine. DOI: 10.1016/S2213-2600(13)70231-

And:

C. U. Koser, J. M. Bryant, J. Becq, M. E. Torok, M. J. Ellington, M. A. Marti-Renom, A. J. Carmichael, J. Parkhill, G. P. Smith and S. J. Peacock. (2013) "Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*." The New England journal of medicine **369**(3): 290-292.

Statement of contribution:

REMoxTB study: I carried out all bioinformatics analyses and interpretation. The study was initiated by SHG and AHD, RD, AP, AAM, CC, IMS, CL, MJB and MH. TDMcH, ALCB, RDH, LW, SM, PvH, and SHG were responsible for microbiological design, conduct, and review of the study. Sequencing was managed by SDB, and SRH gave advice on bioinformatics.

XDR patient study: I carried out all bioinformatics analyses. CUK and SJP initiated the study. Sequencing was carried out at Illumina Ltd. (Cambridge) by JB and GPS. Drug susceptibility testing was carried out Public Health England National Mycobacterium Reference Laboratory (London) except for Clofazimine which was carried out by the Scottish Mycobacteria Reference Laboratory, Edinburgh.

3.1. Introduction

Recurrent tuberculosis infection is defined as a second episode of disease after successful treatment of a previous episode. Recurrence is low globally with the WHO reporting it for 5% of the 6.2 million tuberculosis cases in 2010 (WHO 2012). However, it has been well documented that in high incidence regions such as South Africa, recurrent tuberculosis is more dominant, and is associated with HIV status (Glynn, Murray *et al*. 2010). Recurrence can arise via two routes: relapse of the primary infection that treatment has failed to eradicate, and re-infection with an unrelated exogenous strain.

Until recently re-infection was considered to be rare, as a traditional assumption of tuberculosis epidemiology was that an infection episode is caused by a single strain and that subsequent episodes are caused by re-activation of the endogenous strain (Stead 1967). However there is an increasing appreciation that this is often not the case, and that both mixed infections and exogenous re-infection do frequently occur. This change in thinking is due to the development of genotyping techniques and their application to recurrent tuberculosis disease in a clinical setting, which makes it possible to distinguish if the primary and secondary disease episodes were caused by the same genotype. In a study in India, it was estimated that 88% and 9% of recurrence cases were due to re-infection in HIV positive and negative patients respectively (Narayanan, Swaminathan *et al*. 2010). In South Africa it was noted that the incidence of re-infection was higher than the incidence of new infections, where 77% of recurrence was classed as re-infection (Verver, Warren *et al*. 2005). Similarly mixed infections have also been found to be more common than first thought, with one study in South Africa finding at least two different strains in 19% of patient samples (Warren, Victor *et al*. 2004).

Although these typing techniques have been useful in revealing the possible extent of mixed and re-infections, they can lack resolution as discussed more generally in section 1.4.3. Further to this, mixed samples can be very difficult to detect using traditional techniques, as the signal can be unclear or undetectable if one of the strains is present in too low quantities, or are too similar. This impacts on our understanding of recurrent disease as it would be difficult to disentangle complex scenarios such as

an apparent re-infection which may in reality be a mixed infection followed by endogenous re-activation of one of the strains. The high depth of coverage that can be obtained with whole genome sequencing should allow the detection of mixed infections, and allow us to pick up on these scenarios more accurately.

Here, two studies are presented which both use whole genome sequencing to disentangle the different routes that can result in multiple infections and disease episodes of tuberculosis. The first is based on pairs of samples collected from patients with recurrent disease during a multi-centre clinical trial, REMoxTB. The second is based on a single patient from Addenbrooke's hospital diagnosed with XDR tuberculosis.

3.2. Methods

3.2.1. REMoxTB study

REMoxTB was a phase three clinical trial that aimed to test two four-month moxifloxacin containing regimens compared to standard treatment. 1,931 patients underwent randomised treatment across sites in South Africa, India, Tanzania, Kenya, Thailand, Malaysia, Zambia, China and Mexico (Gillespie, Crook et al. 2014). At the time of the analysis the trial was still ongoing and researchers were blinded to the treatment regimen. The first 50 paired isolates available from participants enrolled in the trial were used: composed of the initial sample upon diagnosis and a post week 17 of treatment sputum sample from patients with relapse or bacteriological failure. Eligible patients were adults diagnosed with previously untreated, drug-sensitive, smear-positive, pulmonary tuberculosis without severe co-morbidities. HIV-positive patients with a CD4-count below 250/µl or those already on antiretroviral treatment were excluded. All subjects providing informed consent were treated for tuberculosis for 26 weeks with one of three different regimens of 4 or 6 months duration that could contain rifampicin, isoniazid, ethambutol, pyrazinamide, moxifloxacin and/or placebo. The total observation period including treatment and follow-up was 18 months.

To distinguish cases due to treatment failure and those resembling recurrent disease, the complete clinical history was reviewed (carried out by A. Bateson, University College London), thereby taking into account all culture results and all clinical

information available. Single isolated positives were also included in order to investigate their clinical relevance, as this is currently unclear. These are cases where a positive culture was followed by at least two negative cultures without re-treatment having been initiated by a physician and the patient remaining symptom free throughout the remainder of follow-up.

Both DNA extraction and MIRU VNTR were performed by R. Hunt and A. Bateson (University College London). MIRU-VNTR typing analyses the number of repetitive DNA sequences at multiple independent genetic loci (ETR-A, B, C, D, E and MIRU-02, 10, 16, 20, 23, 24, 26, 27, 39, 40) as described previously (Supply, Allix *et al*. 2006).

Samples were pair-end sequenced with a read length of 100bp on the Illumina HiSeq platform. The raw sequencing data was mapped to H37Rv and variant calling was carried out as described in the Methods 8.2 and 8.3. Mixed based calls were detected as described in Methods 8.9.

3.2.2. XDR patient study

Sputum specimens taken at the Cambridge University Hospital were processed by laboratory staff at the Cambridge Public Health England Microbiology Laboratory. DNA was extracted by Claudio Köser (University of Cambridge) from one half of a Mycbacterial growth indicator tube (MGIT) culture grown from the first sputum specimen obtained on admission to Cambridge University Hospital. DNA was also extracted from *M. tuberculosis* grown from subculture of the MGIT tube onto a Löwenstein–Jensen (LJ) slope. Library preparation and DNA sequencing (paired-end, 150 bp reads, Illumina MiSeq platform) were performed by Illumina Cambridge Ltd. Mapping and variant calling were carried out as described in Methods 8.2 and 8.3. Mixed base calls were detected as described in Methods 8.9.

3.3. Results – REMoxTB

3.3.1. Overview

Paired samples from 50 patients were sequenced (see Appendix 9.2 for meta-data on pairs). For 96 of the samples (representing 47 patient-pairs plus two singletons where one sample of the pair failed to sequence) an average coverage of 120 fold was obtained, with the remaining four excluded due to poor coverage or contamination with a non-mycobacterial source. Based on the 10,354 variable positions detected, a maximum likelihood phylogeny was built revealing the presence of four of the globally recognized lineages (Gagneux, DeRiemer *et al*. 2006) (Figure 13).

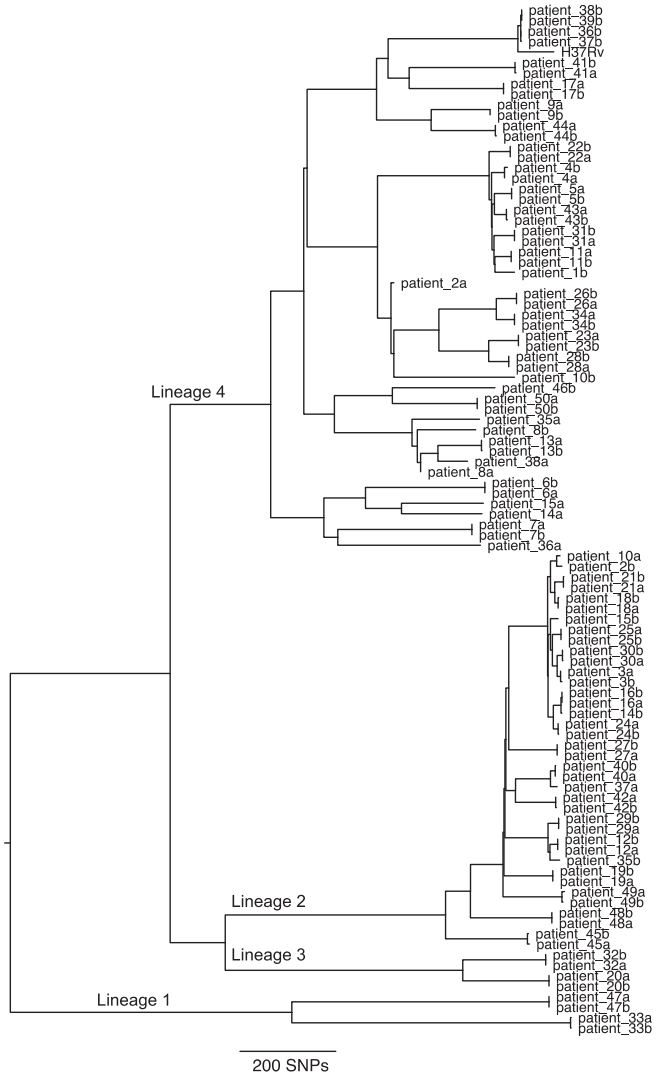


Figure 13 – Maximum likelihood tree of all successfully sequenced isolates in the dataset. Four of the major lineages of the MTBC (Gagneux, DeRiemer *et al*. 2006) are marked. Sample 2a and 8a sit close to internal nodes – they were later found to be a mix of two strains.

Using the observed SNPs between the initial and recurrence strains, cases were defined as relapse (n=33), re-infection (n=3) or mixed infection (n=6) (Figure 14). The rationale behind making these designations are discussed below.

3.3.2. Distinguishing relapse and re-infection

There was a clear distinction between pairs with a low SNP difference (<=6), and those with a high SNP difference (=>1306) (Figure 15). Previously it was observed that within-patient diversity didn't exceed 14 SNPs (Walker, Ip *et al*. 2012), which supports the inference that the low SNP distance pairs represent relapse, and the high SNP distance pairs represent re-infection. Relapse was identified in 33 cases (70% 33/47) with pairs differing by a mean of 0.4 SNPs and the majority (n=27) having no polymorphisms. For three pairs (7%), which all had SNP differences greater than 1306, their recurrence stain was defined as a re-infection. All three involved isolates belonging to different lineages: either the Euro-American or East Asian type. The mean SNP distance between the re-infection pairs was 1355 (Figure 15) which is significantly larger than the mean pairwise distance observed between all isolates in the dataset (972), when compared using the Wilcoxon test (P=0.044).

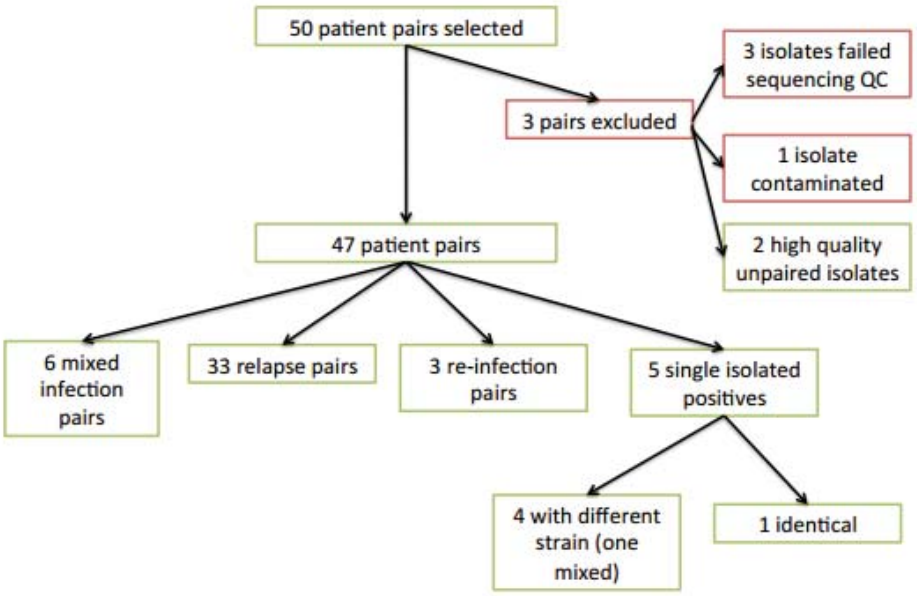


Figure 14 - Summary of sequencing results. Green boxes indicate isolates included in the analysis, red were excluded due to sequencing failure or contamination.

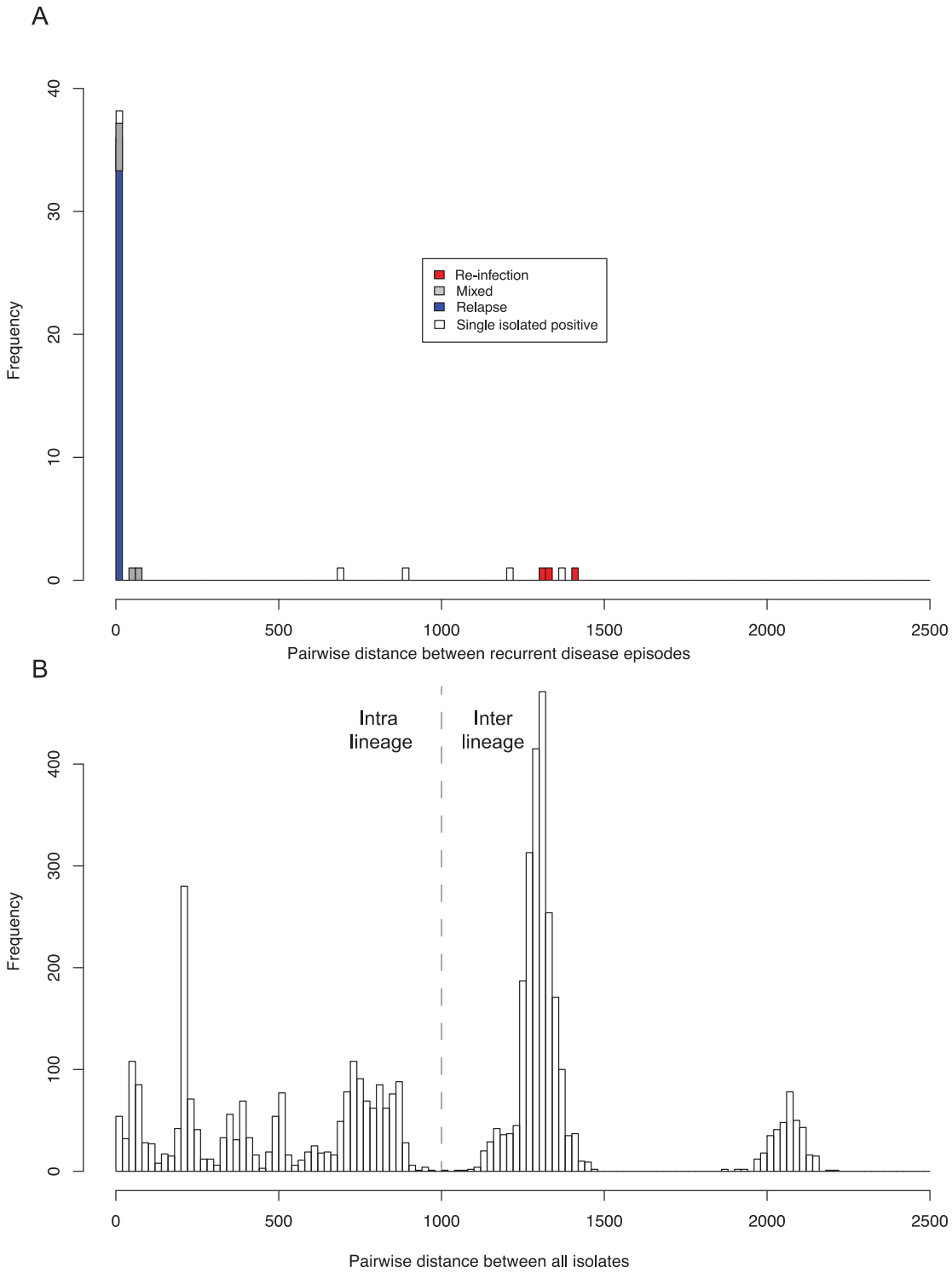


Figure 15 - Histograms of genome-wide pairwise SNP distance between isolates in the dataset. A) Pairwise difference between same patient pairs. B) Pairwise differences between all isolates in dataset, with a dotted line representing the point where the difference represents the distance between two lineages.

3.3.3. Mixed infections

For the majority of samples (n=87) fewer than 40 sites with a mixed base call were identified across the genome, and these were likely to be due to mapping error. In seven sample pairs, however, there were outliers with more than 80 sites (Figure 16) which were manually inspected to look for mixed base calls at lineage defining positions (Stucki, Malla et al. 2012), or where SNPs had been identified in the other isolate of the pair. A total of six patients showed evidence of a mixed infection of which four were mixed in the first sample of the patient pair. They were found to be heterogeneous in positions where a SNP was identified in the second sample, indicating that the initial sample was composed of the strain found only in the secondary isolate plus a sequence from another lineage. Two patient pairs had evidence of two distinct strains only in the secondary isolate (Table 2), one of which was the same strain found in the initial sample, which could be interpreted as relapse and super-infection. An additional sample was also found to have evidence of a mixed population, but was defined clinically as a single isolated positive (see below).

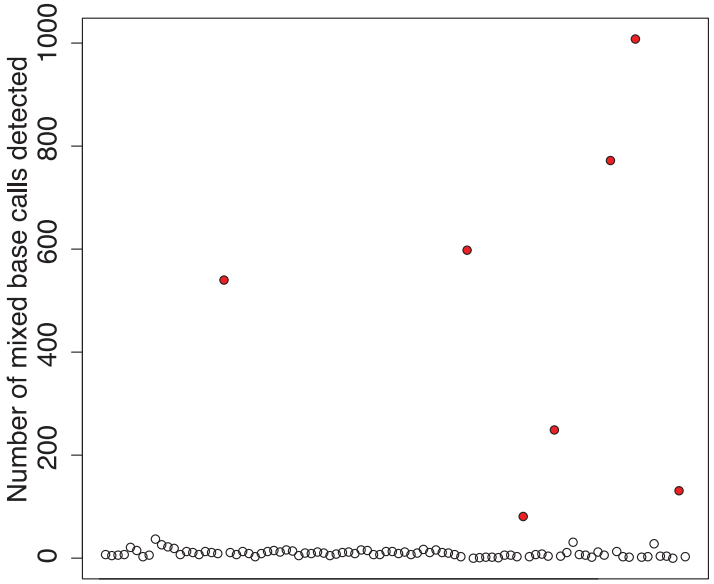


Figure 16 - Number of mixed base calls identified for all isolates in the study. Each dot represents a isolate, arranged in a random order along the x-axis. Red dots represent those identified as mixed.

Table 2 - Proportion of reads matching lineage defining SNPs identified in the mixed infections
Lineage specific SNPs were identified using informative positions previously defined (Stucki, Malla *et al*. 2012) Frequencies represent the proportion of reads that match the base that defines the lineage. * Sample 8a is composed of two Euro-American strains divergent by at least 132 SNPs in a 50% mix ** Sample 42b is composed of a Typical Beijing isolate identical to 42a (95%) plus an Atypical Beijing strain (Schurch, Kremer *et al*. 2011) (5%). Manual inspection of 42b also reveals reads matching the Atypical strain (~2%). The mixed sample from a single isolated positive was excluded.

Patient sample	1. Indo oceanic 3920109 (G->T)	2. East Asian 1834177 (A->C)	3. East African Indian 301341 (C->A)	4. Euro American 3326554 (C->A)	5. West African 1 1377185 (C->G)	6. West African 2 2427828 (C->G)
2a	0	0.16	0	0.84	0	0
2b	0	1	0	0	0	0
8a	0	0	0	1*	0	0
8b	0	0	0	1	0	0
23a	0	0.39	0	0.7	0	0
23b	0	0	0	1	0	0
42a	0	1	0	0	0	0
42b	0	0.96**	0	0	0	0
45a	0	1	0	0	0	0
45b	0	0.93	0	0.08	0	0
50a	0.26	0	0	0.69	0	0
50b	0	0	0	1	0	0

3.3.4. Single isolated positives

Cases were defined clinically as single isolated positives on five occasions. These are incidences where a single sample is found to be sputum positive for *M. tuberculosis*, and in the absence of treatment all subsequent samples are negative. These are usually attributed to lab cross-contamination. Out of the five cases, three of them were with a strain unrelated to the primary case (>500 SNPs), one was mixed and one differed by only three SNPs. The small SNP distance in the latter suggests that this case represents a true relapse and not contamination.

3.3.5. Correlation with MIRU-VNTR data

MIRU-VNTR, one of the most commonly used typing techniques, was carried out on all samples. The three cases identified as re-infection by whole genome sequencing differed by 1-13 loci. Twenty-seven of the relapse cases had an identical MIRU-VNTR type, but five differed by one or more loci. There were six cases identified by

genome sequencing as possible mixed infections but MIRU/VNTR identified four of these as re-infections and two as relapse.

3.4. Results – XDR patient

Two samples were sequenced from a male patient diagnosed with XDR tuberculosis at Addenbrooke's hospital. They were isolated from different culture techniques: first from a MGIT tube, and the second from an LJ slope which are both standard techniques used to select for and grow mycobacteria. When mapped against the *M. tuberculosis* reference genome, a high number of mixed base calls (n=421, raw unfiltered) were called in the MGIT sample but not the slope sample. Manual inspection of these positions revealed that there was an apparent mixture of an isolate that was highly similar or identical to the slope sample, together with another isolate in a ratio of approximately 70:30. The mapping data was filtered for high quality mixed base positions (n=224) and the alleles were separated into two by sorting the alternative alleles for each position into those that matched the slope sample and those that didn't. This enabled the mixture to be separated into "slope-like" and "non-slope-like" as shown Figure 17.

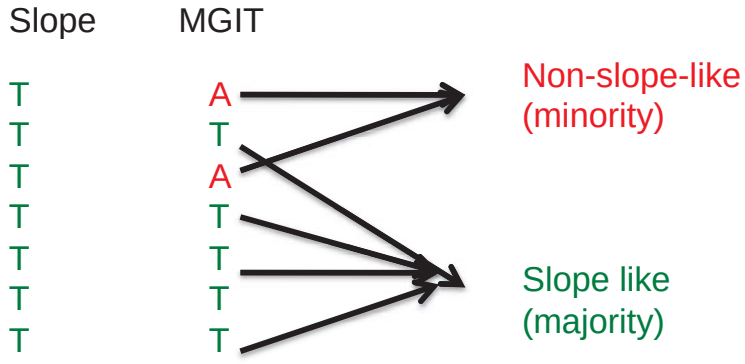


Figure 17 - Rationale used to separate mixed MGIT sample from the XDR patient

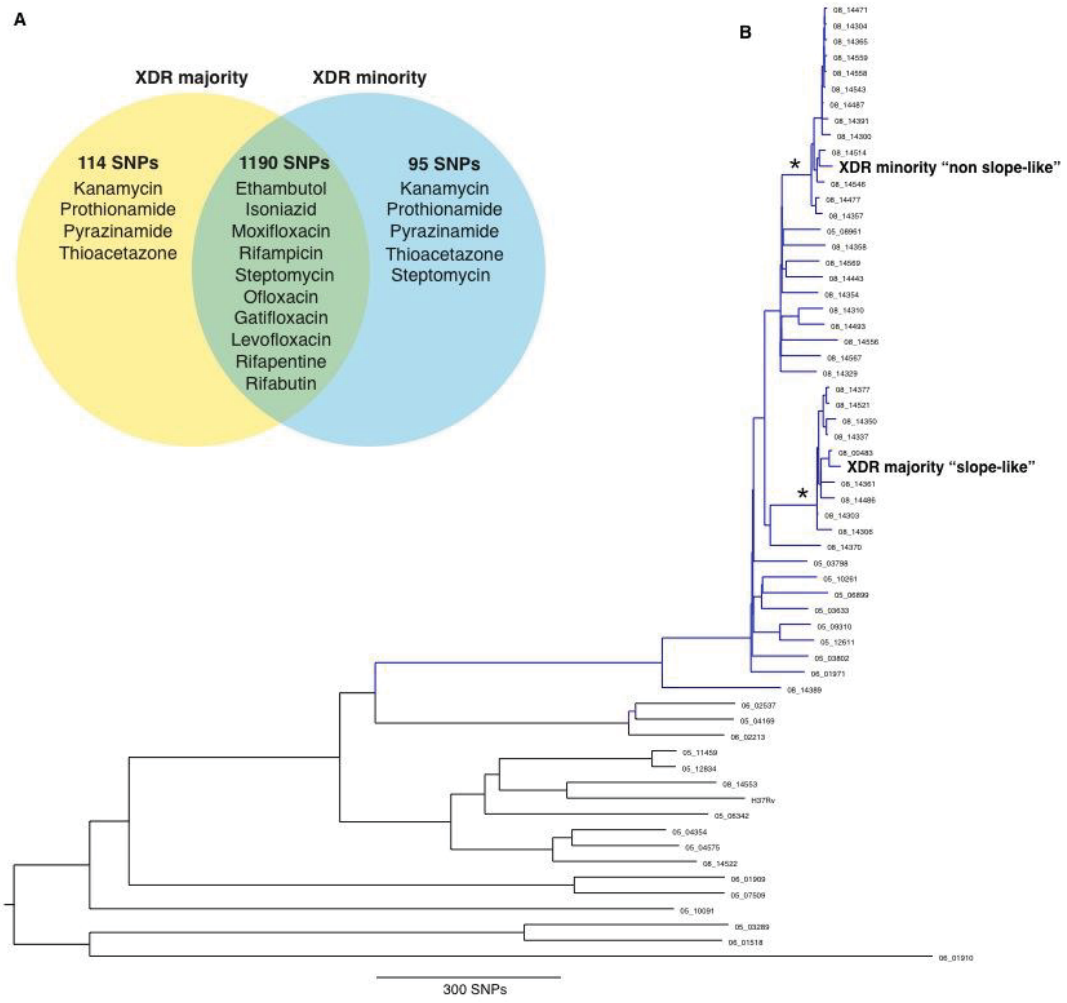


Figure 18 - A mixed extensively drug resistant (XDR) infection. A) Antibiotics with evidence of resistance mutations present in majority and minority strains. The same mutation was present in both strains for ten drugs (green intersect), but different mutations in each strain accounted for resistance to five drugs (yellow and blue). Streptomycin is listed twice as an additional resistance associated variant was found in the XDR minority B) Maximum likelihood tree showing the phylogenetic position of the XDR minority and majority strain in the Beijing (blue) lineage. Contextual strains were from Samara, Russia (Casali, Nikolayevskyy *et al*. 2012). The bootstrap support for the blue clade, and the two clades containing the separated strains (node marked with *) were all 100%

The presence of lineage determining SNPs (Stucki, Malla *et al*. 2012) suggested that they both belonged to the Beijing lineage of *M. tuberculosis*. To place the two strains in context, a phylogeny was built using the Beijing lineage samples from a previous study of tuberculosis in Samara, Russia (Casali, Nikolayevskyy *et al*. 2012). The two strains were found to be paraphyletic (Figure 18b); confirming that this mixture was not a result of diversity generated during an infection.

Drug susceptibility testing concluded that the sample was XDR. However this observed phenotype could be the result of just one of the strains in the mixture, both or a combined result of resistance phenotypes contributed by both strains. For some of the tested antibiotics, resistance mechanisms have been well characterised; for others, very little is known. Possible genes (and in some cases specific codons) associated with resistance to the tested antibiotics were identified through an extensive literature search (carried out by C. Köser). These sequences were then checked for the presence of variants with respect to the H37Rv reference (which is fully susceptible to the drugs of interest). If a possible resistance-causing variant was identified it was called as XDR majority if it matched the variant called in the slope sample (yellow –Figure 18a), XDR minority if it didn't match (blue), and both if present in 100% of the reads (green). This confirmed that both strains in the mixture were XDR, and were found to be resistant to the same antibiotics but were due to different independent mutations in five cases.

3.5. Discussion

A traditional assumption of tuberculosis research is that an infection episode is caused by a single strain and that subsequent episodes are caused by re-activation of the endogenous strain (Stead 1967). However, most tuberculosis clinicians and researchers now appreciate that this can often not be the case, and that a number of different scenarios could be underlying a disease episode. This is particularly true for endemic regions, such as South Africa, where HIV may be a driving force and Eastern Europe, where poor infection control and treatment failure may be resulting in multiple infections. The fact that super-infections (resulting in mixed infections) and new infections (resulting in re-infections) can occur in non-HIV positive individuals suggests that the immune protection conferred by the first infection may not always be strong or durable enough to protect against subsequent infections. Immunity to tuberculosis is poorly understood, but we can speculate that this could be due to either the diversity of the host immune response, or diversity of the pathogen.

There is a possibility that the immune protection conferred by one tuberculosis strain may not extend to more distantly related strains due to differences in their antigenic profile. Hints that this may be the case come from studies of the Bacillus Calmette–

Guérin (BCG) vaccine, a live attenuated form of *M. bovis*. Estimates of its efficacy have varied wildly, ranging from 0 – 90% (Fine 1995), and have been attributed to a variety of factors including host diversity and exposure to NTMs. However, these studies lack knowledge of the prevalence and diversity of circulating *M. tuberculosis* strains, leaving this diversity unaccounted for. Interestingly both animal (Lopez, Aguilar et al. 2003, Tsenova, Harbacheuski et al. 2007) and human (Kremer, van-der-Werf *et al*. 2009) studies have both supported the provocative idea that the highly successful Beijing lineage may represent a BCG vaccination escape variant (Abebe and Bjune 2006). This strain-specific variation in the efficacy of BCG suggests that different lineages of *M. tuberculosis* may confer differential immune protection. Was there any evidence for this in the re-infection cases in this dataset? Unfortunately, three cases are not enough to make any robust conclusions. But it's noteworthy that all the re-infection cases were with strains from a different lineage, and that the SNP difference was significantly larger than would be expected by chance if re-infection were equally likely for all strains in the dataset. In another study focusing on applying whole genome sequencing to transmission chains in Uganda, two re-infection cases were identified. One of these involved two strains from the same lineage (lineage 4), and the other was with strains from two different lineages (lineage 4 and 3) (Clark, Mallard *et al*. 2013). Clearly further studies on larger datasets will be required to address this question, which may have important consequences for vaccine design.

Future studies will need to use whole genome sequencing to accurately distinguish the scenarios of relapse, re-infection and mixed infections. This is reflected by the fact that this study found that 11/47 cases came to different conclusions than those using the MIRU-VNTR data. In the context of a clinical trial, this means that 6/33 cases were misclassified as relapse: the primary end-point in a clinical trial. This high level of misclassification could also impact on our understanding of the prevalence of these scenarios. Previous studies have used a cut-off of greater than one locus to conclude re-infection (Narayanan, Swaminathan et al. 2010, Martin, Herranz et al. 2011), and this would have resulted in the misclassification of two of the relapse pairs (differing by 2 and 3 loci), which means that re-infection may have been over estimated in these cases.

One major limitation of all genotyping techniques, including whole genome sequencing when attempting to classify recurrent disease, is that an apparent relapse may be due to re-infection with a closely related strain (from a family member for example). It's not known how often this occurs, and would be impossible to estimate using the approach described here. In order to accurately quantify the rates of these two processes, future analyses may need to incorporate modeling approaches and epidemiological information collected from patients.

The detection of mixed infections is important for individual patient management in addition to increasing our understanding of tuberculosis epidemiology. The XDR case described here demonstrates that different infecting populations in the same patient can have different resistance profiles, and that whole genome sequencing provided clarity in this respect. Mixed infections are expected to be more difficult to treat and more likely to lead to acquired resistance (Cohen, van Helden *et al*. 2012). Furthermore, miss-identification of a mixed infection could lead to errors during epidemiological investigation, when failure to detect both strains in the index case could lead to failure to define a transmission event to secondary cases. On the population level, mathematical models predict that a high preponderance of mixed infections will lead to the survival of less fit strains, which will persist longer than they would in the absence of mixed infections (Cohen, van Helden *et al*. 2012). More complex models also predict that if mixed infections were common, control interventions that target latent infection (such as isoniazid preventative therapy) would be more likely to lead to the emergence of drug resistant strains (Colijn, Cohen *et al*. 2009).

This study identified six mixed infections, despite the bacteriological methods being orientated towards the isolation of a single strain, suggesting that this is an underestimate of the real burden. Previous estimates of the prevalence of mixed infections were based on genotyping techniques, and were often limited to one sample per patient. A study in Georgia, found that out of the 26 mixed infection cases that were identified using genotyping of multiple samples, all or 14 (RFLP typing or PCR respectively) of them would have been missed entirely based the analysis of a single pre-treatment isolate (Shamputa, Jugheli *et al*. 2006). This demonstrates that to really understand the prevalence of mixed infections in different settings, whole genome

sequencing will need to be carried out on multiple colonies or non-colony-purified cultures of multiple samples. The XDR study also highlights the possibility that laboratory handling of samples may result in selection for one of the strains, as the mixture was only identified in the MGIT sample, and not the LJ slope sample.

Of the five cases identified as single isolated positives in this study, four were likely due to cross contamination and one provides evidence for the first time, that positive cultures originating from the patient's own infection may be cultured and the patient's infection resolved without further treatment. Cross contamination is a well-recognised challenge in myco-bacteriology laboratories, accounting for up to 3.9% of samples (Glynn, Yates *et al*. 2004). It usually occurs in less than 1% of positive samples with more than half of laboratories achieving a rate of less than 2.5% (Ruddy, McHugh *et al*. 2002). In the clinical trial setting there is a need to ensure that adequate molecular methods are in place to identify the origin of single isolated positive samples correctly.

The ability to accurately distinguish relapse, re-infection and mixed infections is of critical importance for an understanding of tuberculosis epidemiology, determining end points in clinical trials and for patient management. This study provides a proof of principle demonstrating that whole genome sequencing can distinguish these different scenarios unequivocally. Larger scale studies will now be required in order to quantify these processes in different geographical, clinical and social contexts.

4. Diversity of the PE and PPE gene families from *Mycobacterium tuberculosis*

Part of this analysis has been published in:

J. M. Bryant, S. R. Harris, J. Parkhill, R. Dawson, A. H. Diacon, P. van Helden, A. Pym, A. A. Mahayiddin, C. Chuchottaworn, I. M. Sanne, C. Louw, M. J. Boeree, M. Hoelscher, T. D. McHugh, A. L. C. Bateson, R. D. Hunt, S. Mwaigwisya, L. Wright, S. H. Gillespie and S. D. Bentley (2013). "Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study." The Lancet Respiratory Medicine. DOI: 10.1016/S2213-2600(13)70231-

Statement of contribution

I performed all bioinformatic analyses and interpretation. PacBio sequencing and assembly was performed by Paul Coupland (WT Sanger Institute).

4.1. Introduction

The first genome sequence of *M. tuberculosis* (Cole, Brosch *et al*. 1998) revealed two novel gene families characterized by conserved Proline-Glutamate (PE) and Proline-Proline-Glutamate (PPE) residues at their N termini. These genes have been of significant interest due to their number, diversity and cell surface localization. Despite this, an understanding of these genes remains elusive and a clear function is yet to be determined.

A total of 168 PE and PPE genes have been described for *M. tuberculosis*, which make up almost 10% of its genomic coding capacity. The families are unique to Mycobacteria, found in varying numbers across the genus and are commonly cited to be particularly abundant in the more pathogenic species (Sampson 2011). However this observation may be merely anecdotal, as it appears that the number of genes seems to be more related to the traditional genus division into slow and rapid growers rather than their propensity to cause disease (Table 3).

Table 3 - Numbers of PE and PPE genes annotated in the genomes of Mycobacterial species. Genome annotations acquired from NCBI (NC_008596, NC_008726, NC_009338, NC_009077, NC_008703, NC_008146, NC_010394, NC_010604, NC_005916, NC_000962,NC_008595, NC_002944, NC_002677). * 27 PE/PPE pseudogenes not counted.

Species	PE	PPE	Pathogenic to humans?	Growth phenotype
M. smegmatis	6	4	No	Rapid
M. vanbaalenii	1	22	No	Rapid
M. gilvium	2	11	No	Rapid
M. sp JLS	2	12	No	Rapid
M. sp KMS	3	13	No	Rapid
M. sp MCS	2	11	No	Rapid
M. abscessus	3	6	Yes- opportunistic	Rapid
M. marinum	170	105	Yes- opportunistic	Slow
M. ulcerans	70	46	Yes- opportunistic	Slow
M. tuberculosis	99	69	Yes- obligate	Slow
M. avium 104	7	35	Yes- opportunistic	Slow
M. avium paraTB K10	6	36	Yes- opportunistic	Slow
M. leprae	1*	2*	Yes- obligate	Slow

The N termini of these proteins, which contain the PE or PPE motifs, are highly conserved in comparison to their C termini, which vary in both sequence and length (Cole, Brosch *et al*. 1998). This variable C terminus is either completely unique or contains specific repeat sequences allowing the families to be further subdivided (Figure 19). The largest subfamily consists of the PE-PGRS genes, which are typically long, extremely GC rich and full of tandem repeats. Other members have a simpler arrangement, and may only be composed of the conserved PE or PPE N termini domain. Only one crystal structure has been solved (Strong, Sawaya *et al*. 2006), which revealed that PE25 and PPE41 (which are found adjacent in the *M. tuberculosis* genome) form a 1:1 heterodimeric complex. It is unknown how representative this complex is of the proteins, but it is interesting that 18 other contiguous pairs of PE and PPE genes are found across the *M. tuberculosis* genome, suggesting that they too may form heterodimers as proteins.

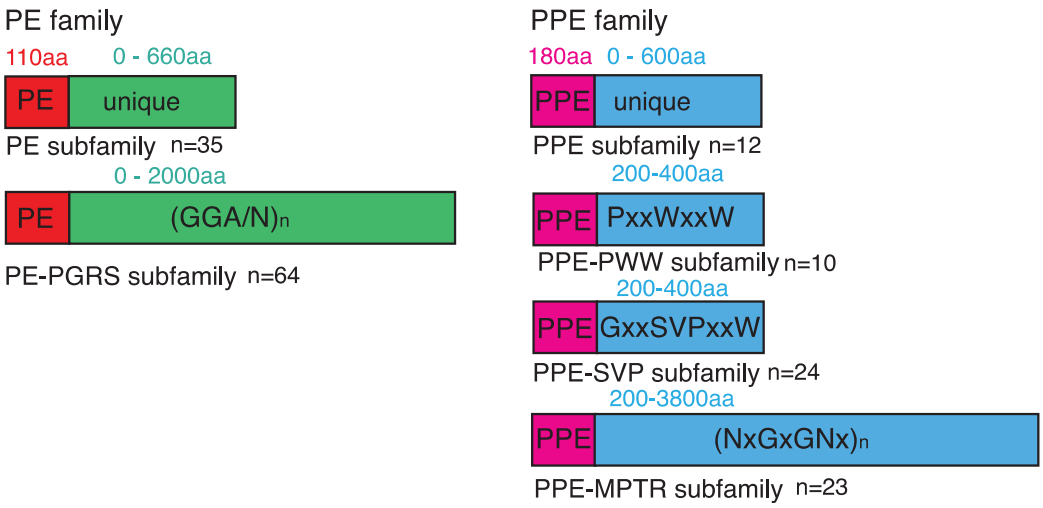


Figure 19 - Subfamilies and structure of the PE and PPE genes. The families are characterised by a conserved N terminal domain (red and pink) with variable C termini (green and blue). The amino acid motifs found in the C termini are stated with n indicating that they are tandem repeated. The approximate lengths of these regions in amino acids are indicated. PGRS = polymorphic GC-rich-repetitive sequence, MPTR = major polymorphic tandem repeat. Diagram adapted from Pittius *et al* (Gey van Pittius, Sampson *et al*. 2006).

Our functional knowledge of the PE and PPE genes remains poor, and is limited to a number of disparate observations and studies on individual genes. One of the first observations was that many of the members are found in close proximity to the

ESAT-6 genes, suggesting they may have co-expanded via duplication (Gey van Pittius, Sampson *et al*. 2006). These genes encode proteins that form part of the type VII secretion apparatus responsible for the secretion of the immuno-potent and antigenic ESAT-6 proteins (Gey Van Pittius, Gamieldien *et al*. 2001) and experimental evidence from *M. marinum* suggests that it is also likely to mediate the secretion of the PE/PPE proteins (Abdallah, Verboom *et al*. 2009). A second key observation is that many of the proteins are thought to be cell surface associated or exported, as demonstrated experimentally for 22 of the members (reviewed by Sampson 2011). In addition many members are expressed specifically during macrophage invasion (Rachman, Strong *et al*. 2006) or granuloma formation (Ramakrishnan, Federspiel *et al*. 2000). Collectively, these findings suggest that in *M. tuberculosis* at least, the PE and PPE genes are intimately involved with pathogenicity and/or interactions with the human immune system. However, we are yet to determine a more specific role.

There is a great deal of evidence supporting the observation that there is a high level of diversity between isolates, for both individual genes (Talarico, Cave et al. 2005, Talarico, Zhang et al. 2008) and the gene families as a whole (McEvoy, Cloete *et al*. 2012). The PE-PGRS and PPE-MPTR subfamilies are thought to contribute most of this diversity, mainly through mutations in their long repeat-rich C terminal domains (Sampson 2011) (Figure 19). This has led to speculation that these genes may contribute to evasion of host immune responses through antigenic diversification (Cole, Brosch et al. 1998, Karboul, Mazza et al. 2008). However, there has been no direct evidence demonstrating within patient antigenic variation (Sampson 2011). One of the reasons our understanding of these genes is very much incomplete is that they are hard to both sequence and to analyse, due to their high GC content (up to ~85%) and their repetitive nature. Although the presence of diversity through SNPs, insertions/deletions and intra-genomic recombination has been observed by many studies, the parameters of this diversity have not been quantified. It is unknown at what rate it is generated and similarly at what level: whether at the lineage, cluster or within patient level. Understanding these processes has important implications for our understanding of these gene families.

Here, two analyses of these gene families are presented. The first utilises sequences generated for the ReMoxTB project (Chapter 3) where pairs of isolates were collected from patients with multiple disease episodes. By looking at the PE/PPE genes in relapse cases, this study aimed to determine if there is variability generated within patients. The second analysis utilised sequence data from a study on the Beijing lineage in Samara, Russia (Casali, Nikolayevskyy *et al*. 2014). The Beijing clone is an important lineage which is thought to have recently spread across Asia and Eastern Europe, and has gained much interest due to its rapid global dissemination and high drug resistance. The aim of this work was to determine how much variation had been generated in the PE/PPE gene families and relate this back to the phylogeny.

4.2. Methods

Previously high-throughput sequence analysis of the PE and PPE genes was impossible due to inherent GC bias in the sequencing technology. This was overcome by the sequencing development team (WT Sanger Institute) who found that library preparation with the Kapa Hifi enzyme greatly improved coverage in GC rich regions (Quail, Otto *et al*. 2011). For this reason all the library preparation for Illumina sequencing in these analyses was carried out with Kapa Hifi (Kapa Biosystems, MA USA).

Sequencing data from 48 isolates were used for the first part of this study, which are composed of 24 pairs of isolates from baseline and relapse disease episodes collected during the ReMoxTB clinical trial. Velvet (Zerbino and Birney 2008) was used to *de novo* assemble the reads with scaffolding enabled (Methods 8.4). Raw reads were mapped back to the assembly to mask or correct possible assembly errors. Sequences of the 160 PE and PPE genes annotated in H37Rv, were extracted from the assemblies using an in-house script that uses a simulated PCR approach where upstream and downstream ‘primer’ sequences are specified. Alignments were made using Muscle (Edgar 2004) for each gene for the 48 isolates. SNPs, insertions and deletions were identified between the pairs using a custom script. The few differences identified were manually checked assessed using raw mapping data.

For the second part of the study, the sequencing data from 186 isolates available from a study in Samara, Russia were used (Casali, Nikolayevskyy *et al*. 2014), which all belong to the Beijing lineage. PacBio sequencing was carried out on one of these samples by Paul Coupland in the sequencing development team (WT Sanger Institute). The reads were *de novo* assembled and corrected using software provided as part of the SMRT Analysis software package (HGAP and Quiver - https://github.com/ PacificBiosciences/SMRT-Analysis). The short-read Illumina data from the same sample was then mapped onto the reference (Methods 8.2), in order to correct any possible single-base assembly errors, of which there were only ten. Members of the PE and PPE gene families were identified using a simulated PCR approach, and manually annotated. Low complexity regions were identified using DustMasker (Morgulis, Gertz *et al*. 2006) on default settings. The raw sequencing data from the 186 isolates were then mapped onto this final reference using GATK to realign insertions and deletions (Methods 8.2). Variants were called and trees built as described in Methods 8.3 and 8.6.

4.3. Results

4.3.1. Diversity of PE/PPE genes in relapse cases

Overall 86% of genes (82% PPE, 97% PE, 84% PE-PGRS) were assembled which enabled 82% of comparisons to be made between the 24 pairs. A minority of genes could not be assembled at all (see Appendix 9.3). Many of these can be attributed to common deletions such as PPE57, PPE58 and PPE59 which are found in the RD6 region and are flanked by IS sequences (Brosch 2002). Similarly PPE38 is a hotspot for IS6110 integration (McEvoy, Cloete *et al*. 2012); PPE54 contains tandem repeats and PPE46-47 are near identical so could not be assembled.

When comparing the assembled genes between the ReMoxTB pairs, no SNPs could be identified. A few single base pair deletions were detected, however, these were considered untrustworthy as they were found in homopolymeric tracts, which are susceptible to a common assembly error caused by collapsed repeats. In addition they were not supported by mapping data when mapped to the H37Rv reference.

4.3.2. Diversity of PE/PPE genes across the Beijing lineage

4.3.2.1. PacBio reference sequence

Although an assembly approach has its advantages, it can be very labour intensive and not suitable for high throughput analyses. Mapping is quick and requires less manual curation, as it is not prone to artifacts that are the result of miss-assembly. Mapping does however, suffer from the problems of mismapping especially when the reference is too genetically disparate from the isolate in question. Thus in order to use a mapping approach to investigate PE/PPE diversity across the Beijing lineage, a suitable high quality reference is required, as genetically similar as possible to the samples under investigation. PacBio sequencing was carried out on a sample from the Samaran collection, 2535G, which was chosen on the basis of its high coverage (via Illumina sequencing) and central phylogenetic position within the lineage. An assembly of the raw PacBio reads resulted in 11 contiguous sequences.

Members of the PE/PPE genes on this sequence were identified using an *in silico* PCR approach using primers designed from the H37Rv sequence. In total, 157 of the 159 genes found in H37Rv were successfully annotated. All had assembled completely, except for PE_PGRS6 which was situated on a contig boundary. Only small fragments of PE_PGRS9, wag22 and PE_PGRS43 were found at the ends of contigs, and so were excluded from further analyses. A number of genes were identified as pseudogenes due to truncation or frameshifts outlined in Table 4.

Table 4 - Pseudogenes identified in reference sequence, 2535G. *Undetermined due to the requirement of high quality assemblies to assess whether this IS element is present at this position in all the strains.

Gene	Type of mutation	Shared by all isolates?
PE_PGRS13	frameshift	All
PE_PGRS57	frameshift	just S535G
PPE34	truncation due to IS element	Undetermined*
PPE38	truncation due to IS element	Undetermined*
PPE39	truncation due to IS element	Undetermined*
PPE56	frameshift	All
PPE57	frameshift	All
PPE58	frameshift	All
PPE66	stop codon	All

4.3.2.2. Rate of SNP accumulation in PE/PPE genes

186 isolates from the Samaran collection were mapped to the reference sequence, and an average of 95.1% of the base calls in the PE and PPE genes passed quality filters. The resultant phylogenetic tree (Figure 20) revealed a clonal structure as described previously (Casali, Nikolayevskyy *et al.* 2012). Two outlying isolates (non-Beijing but belong to the East Asian clade) were excluded from further analysis but were used as out-groups to correctly root the tree.

In total, 3931 sites across the genome were found to have a high quality SNP, of which 6.3% were within the PE and PPE genes (n=249). This is extremely close to what would be expected by chance (the PE and PPE genes make up 6.2% of the genome), suggesting that there is not an excess number of SNPs in these genes. When the number of PE/PPE vs. non-PE/PPE SNPs per branch was plotted there was a similar relationship (Figure 21), where the proportion of PE/PPE SNPs was found to be 7%.

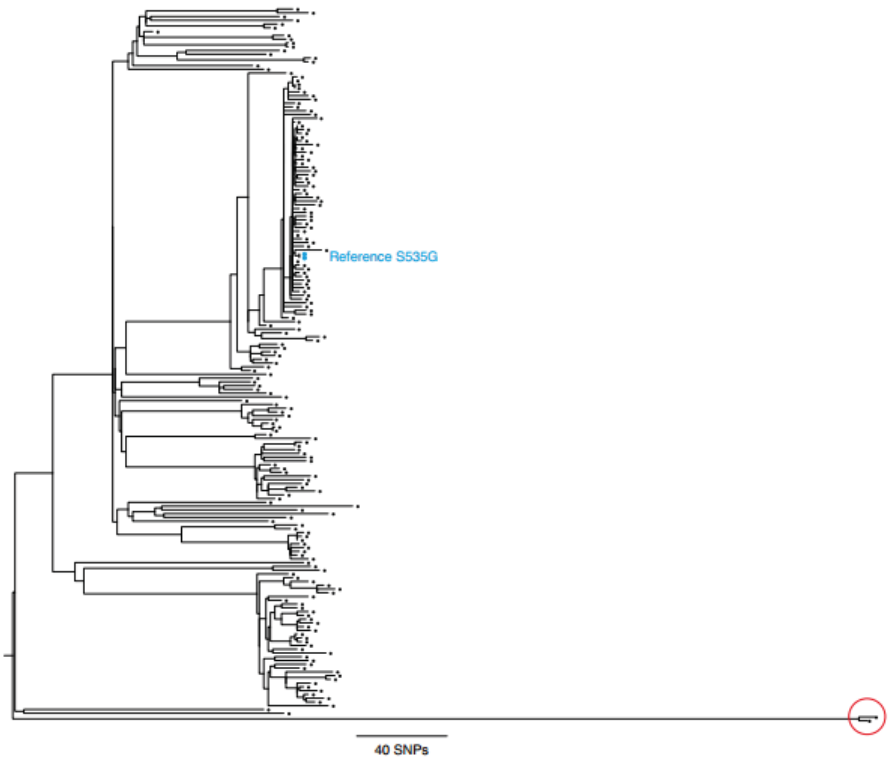


Figure 20. Maximum likelihood phylogeny of Beijing isolates from Samara. Isolates circled in red were excluded from further analyses but were used as an outgroup. The reference strain (both PacBio assembly and the mapping of original Illumina reads) are indicated in blue.

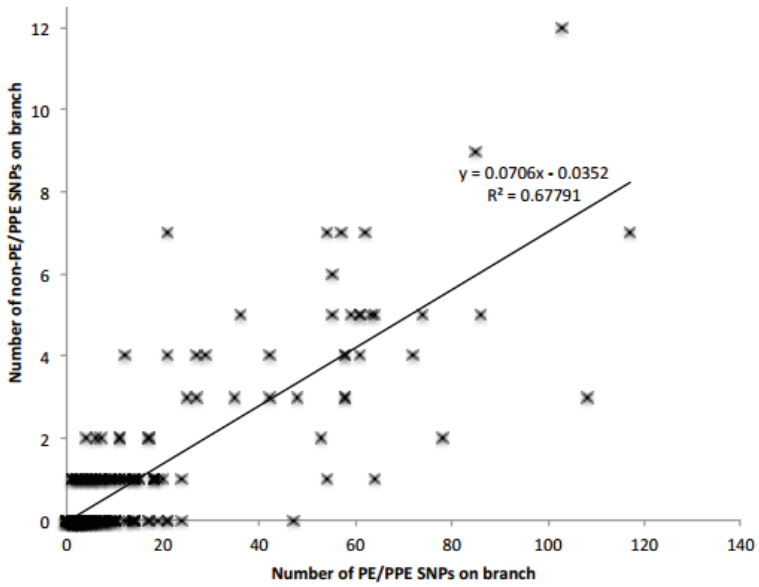


Figure 21 – Plot of number of SNPs found in PE/PPE genes vs. rest of genome. Each point represents a branch on the maximum likelihood phylogeny. Linear regression model p-value: < 2.2e-16.

The total dN/dS of all the PE/PPE SNPs was 0.77, which is slightly higher than the value calculated for the rest of the genome (0.66). When using the raw numbers of non-synonymous and synonymous sites, this was not significantly different using Fishers exact test (P=0.35). Overall this SNP analysis suggests that within this lineage there is no difference in the rate or type of SNP accumulation in this gene family in comparison to the rest of the genome.

4.3.2.3. Insertions and deletions

Using a mapping approach, 134 insertions and deletions were detected within the PE/PPE genes, which were reconstructed onto the SNP-based phylogeny. Events that were found to be homoplasic (n=26), were inspected manually. 17 of these were found within two tandem homopolymeric tracts of PPE13:

GTGCCCCCCCCCCCAAAAAAAAGTA,

Homopolymeric tracts are highly susceptible to frameshift mutations, which were found to occur within these tracts 80 times, making these events extremely frequent. This region is positioned close to the C terminus of the gene (within last 12 amino acids), with a stop codon at a similar position in each alternative frame, making it less likely this frameshift could be deleterious. The other nine homoplasic deletions all occur in a small number of strains, close to the tips of the tree, and appear to be the

result of some variants passing quality filters in some strains but not in others. All these homoplasic variants (including the PPE13 frameshifts) were excluded from further analyses as they might obscure interpretation. An additional six insertions found close together within one strain were also excluded, which were likely due to a combination of low coverage and low level contamination from another bacterium. This left a total of 102 high quality insertions and deletions to interpret.

Overall when correcting for the number of sites, approximately four fold more insertions and deletions were detected in the PE/PPE genes than the rest of the genome (Figure 22). The statistical significance of this difference could not be assessed here but could possibly be investigated with more sophisticated evolutionary analyses implemented in a package such as BEAST ((Drummond and Rambaut 2007).

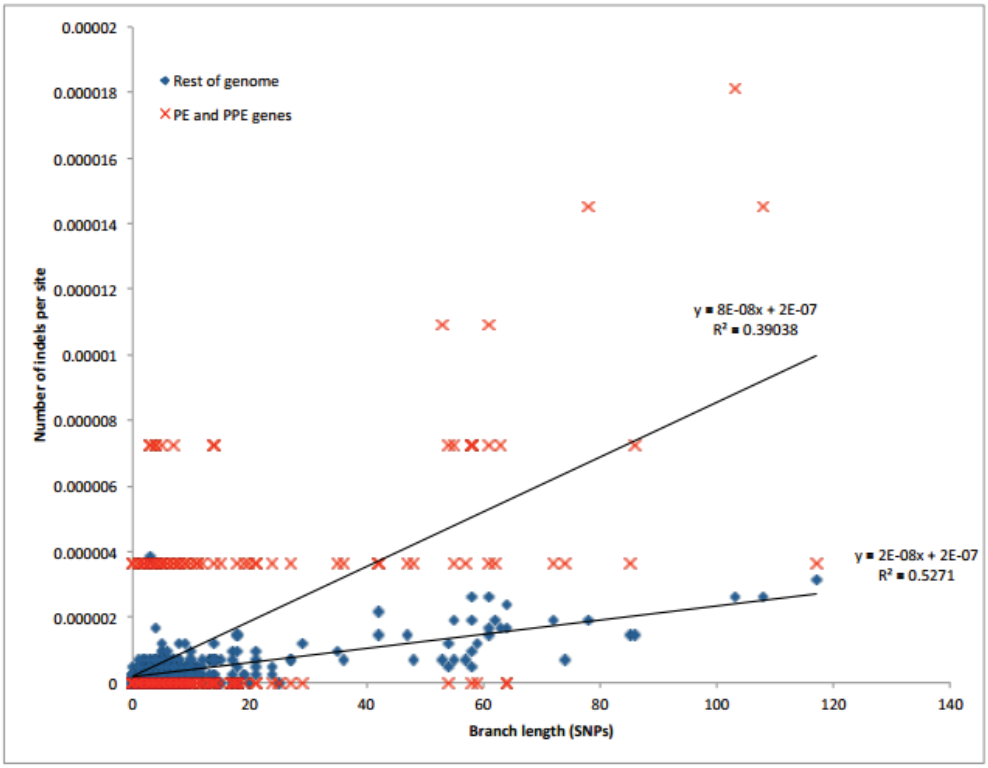


Figure 22 – Number of insertions and deletions per branch found in PE and PPE genes (red) and rest of genome (blue) corrected for number of sites. The PE/PPE genes had a total size of 275650bp, and the rest of the genome was 4176827bp. Linear regression R squared values and equations are shown.

Insertions and deletions were found in the PE-PGRS and PPE subfamilies, but were completely absent from the PE genes (Figure 23). When considering the total length

of these gene families, the number of PE-PGRS variants (0.02 per nucleotide) outweighs those found in PPE genes (0.002), by 10 fold. The majority of the variants (n=58) maintained the frame of the gene (length was a multiple of 3), so are less likely to be deleterious. Interestingly, the majority of these in-frame variants were found in the PE-PGRS subfamily (**Table 5**), and this was significantly more than would be expected by chance (Fishers exact test P = 0.0002). By chance, a third of variants would be expected to be in-frame; in the PE-PGRS genes the opposite was found, where 71% maintained the frame.

Table 5 – Number of insertions and deletions found in the PE and PPE genes.

	PE_PGRS subfamily	PPE family
In frame	54	4
Out of frame	22	14
Whole gene deletions	3	7

2 of the PPE and 1 of the PE_PGRS whole gene deletions were a single event.

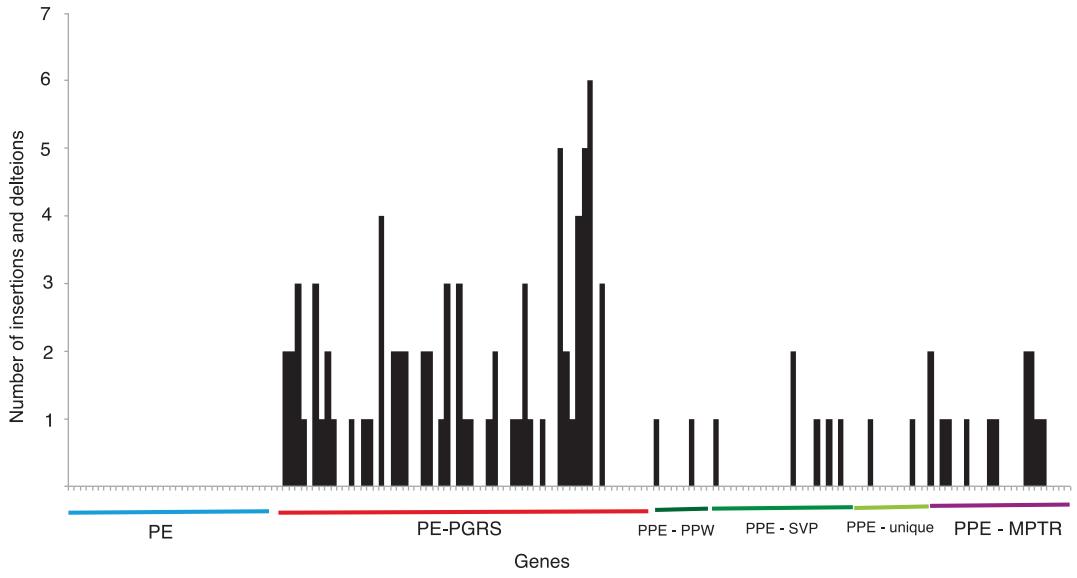


Figure 23 – Numbers of insertions and deletions found in the genes – excluding homoplasies and PPE13 homopolymer frameshifts.

Next, in order to investigate the sequence context of these insertions and deletions, DustMasker (Morgulis, Gertz *et al*. 2006) was used to predict the co-ordinates of low complexity or repetitive sequences. For the in-frame variants, the majority (51/58) were found in these repetitive regions, which was significantly more than the out-of-frame variants (17/36; Fishers exact test P value <0.001).

4.3.2.4. Intra-genomic recombination

There is plenty of anecdotal evidence supporting the occurrence of intra-genomic recombination between PE and PPE genes (Liu, Gutacker et al. 2006, Karboul, Mazza et al. 2008, McEvoy, van Helden et al. 2009), however these events are yet to have been placed in a phylogenetic context. In order to identify such events within the Beijing lineage, SNPs were identified that occurred within 200bp of one another within the PE and PPE genes that were reconstructed onto the same phylogenetic branch. After manual inspection to exclude any events that were likely due to mapping error, seven events were identified and the contextual sequence extracted in order to look for a possible donor (Table 6). An exact match in another gene in the same genome was found for three of the sequences, two of them in a highly similar adjacent gene. None of the donor genes had a reciprocal event suggesting that gene conversion without crossover was the most likely mechanism. All three occurred on terminal branches, perhaps suggesting that these events aren't readily fixed in the population. The underlying basis for the other four events is unknown, but they could possibly represent the normal *de novo* generation of SNPs, inter-genomic recombination with another genome or mapping error.

Table 6 - Possible intra-genomic recombination events identified. These events represent the occurrence of more than one SNP within 200bp of one another on the same phylogenetic branch. For three of them a potential donor was identified.

Branch (node ID->node ID)	No. of SNPs	Gene	Donor identified?
17->ERR230004	2	PPE5	
293->ERR227999	2	PPE47	Exact match in PPE46
293->ERR227999	4	PE-PGRS50	Exact match in PE-PGRS7
12->ERR228068	2	PPE56	
355->ERR227984	4	PPE56	3/4 found in PPE55
240->ERR227996	3	PE-PGRS55	
root->42	2	PE-PGRS22	

4.4. Discussion

Although it has long been appreciated that there is a high level of genetic diversity in the PE and PPE genes, this evidence has been mainly anecdotal, with a few notable exceptions where studies have attempted to quantify the diversity over sub-sets of the family (McEvoy, Cloete et al. 2012, Copin, Coscolla et al. 2014). This work represents one of the most comprehensive analyses of the two families to date, over the most samples and the most genes. This has been achieved in part because of advances in sequencing technologies developed here at the WT Sanger Institute, which significantly reduced the effects of GC bias. However, despite this the data still presented methodological difficulties, meaning that the analysis was confined to short evolutionary distances and lacked a species-wide view. However, this did allow a more focused approach where the diversity could be quantified over two specific evolutionary scales.

The first study aimed to investigate diversity generated on the within-patient scale, by comparing two samples from two disease episodes. A striking and singular conclusion of this work was that there was none; there were no genetic differences detected within these genes between disease episodes. This supports previous observations made using RFLP analysis of PE-PGRS genes within patients, where no diversity in gene structure was detected (Richardson, van der Spuy *et al*. 2004). This result strongly suggests that these genes are not a source of antigenic diversity within patients. However this analysis was limited to around the 80% of the genes that could be assembled, so there could be missing variation generated in the remaining 20%. For this reason, in the second study a mapping rather than *de novo* assembly approach was used.

Using a mapping approach, the genetic diversity was captured across more genes and more isolates, as the method has the advantage of being automated without the possibility of miss-assembly. This enabled the investigation of PE and PPE gene diversity within a phylogenetic context across the Beijing lineage. It was found that the rate of SNP accumulation in these genes, and the corresponding dN/dS was indistinguishable from that of the rest of the genome. This suggests that any unusual levels of diversity observed in the PE and PPE genes are not being generated by *de*

novo SNPs, at least on this evolutionary scale. This is counter to previous observations that the PE and PPE genes had 3 fold more non-synonymous variants than the rest of the genome (McEvoy, Cloete *et al*. 2012). However, the variation in that study wasn't put in a phylogenetic context, which meant that they could not distinguish *de novo* SNPs from those introduced via recombination. In this study, there was strong evidence supporting three intra-genomic gene conversion events, which over larger evolutionary distances may be more numerous and a greater contributor to diversity, and thus dN/dS. A species-wide analysis will be required to investigate the contribution of intra-genomic gene conversion to diversity over larger evolutionary distances.

Instead of SNPs, an excess of insertions and deletions were detected: particularly within the PE-PGRS subfamily. These were dominated by in-frame variants located within repetitive or low complexity regions. The excess of in-frame variants was higher than would be expected by chance, suggesting that either there is a strong selection pressure to retain function in these genes, or the arrangement of repeats (in multiples of three bases) promotes events which are in-frame. The repetitive regions of the PE-PGRS genes are dominated by Gly-Ala repeats, which appear to be to be extremely similar to those found in the Epstein-Barr virus nuclear antigens (Cole, Brosch *et al*. 1998). In EBV, these repeats are involved in immune interference of cytotoxic T cells, so it is speculated that they may play a similar role in *M. tuberculosis*. In support of this, it's been found that the PE region alone is able to elicit a cell-mediated protective immune response, but when the PGRS domain is included this effect is lost (Delogu and Brennan 2001). If these genes were involved in immune interference, it's unclear what role if any a high level of genetic diversity would have. A recent study found few predicted T cell epitopes in the variable C termini of PE-PGRS genes (Copin, Coscolla *et al*. 2014), suggesting that this diversity is independent of T cell recognition. It's possible that the high level of diversity generated could be a side effect of the repeat structure required for immune interference, and that most of the in-frame insertions and deletions generated are neutral. However, even if the inter-genomic variation has little apparent function, it cannot be avoided that on the intra-genomic level these gene families are extremely diverse and that this diversity has been generated and maintained in *M. tuberculosis*.

Despite the extensive work carried out on the PE and PPE genes, they remain enigmatic, and yet more genetic diversity and experimental studies are probably required to gain further ground. However the work presented here has provided greater clarity on the underlying processes that are generating diversity within these genes. Furthermore, many in the field have been concerned that current mapping approaches may be missing a significant amount of diversity generated during transmission or outbreaks by excluding the PE and PPE genes. This work confirms that is unlikely to be the case as little variation seems to be generated over those time scales.

5. Transmission of *Mycobacterium abscessus* within a cystic fibrosis clinic

Includes work published in:

J. M. Bryant, D. M. Grogono, D. Greaves, J. Foweraker, I. Roddick, T. Inns, M. Reacher, C. S. Haworth, M. D. Curran, S. R. Harris, S. J. Peacock, J. Parkhill and R. A. Floto . (2013). "Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study." Lancet **381**:1551-60.

Statement of contribution

I carried out all bioinformatic analyses. DMG carried out collection of clinical data and sample preparation and DNA extraction. RAF, JP and SJP conceived the study. DMG, IR, TI and MR carried out epidemiological analysis.

5.1. Introduction

Although thought to be primarily an environmental bacterium, *M. abscessus* can also infect humans via wounds or as a respiratory pathogen. In particular, it is a significant cause for concern in cystic fibrosis patients, where infections are chronic and extremely difficult to treat due to its naturally high level of antibiotic resistance.

Worryingly, cases are thought to be on the rise, as observed in Taiwan, United States and Australia (Lai, Tan et al. 2010, Prevots, Shaw et al. 2010, Thomson 2010). The reason for this rise is not known, although possible explanations include: the increased use of intravenous antibiotics which by removing the resident microbiota may allow increased colonisation of Mycobacteria (Torrens, Dawkins *et al*. 1998); impairment of host anti-mycobacterial immunity through autophagy inhibition by chronic azithromycin therapy (Renna, Schaffner *et al*. 2011); patient-patient transmission; and increased surveillance.

Extensive person-to-person transmission of other cystic fibrosis pathogens, such as *Pseudomonas aeruginosa*, has long been demonstrated. For *M. abscessus* however, this was thought to be impossible or extremely rare; as previous small-scale studies had failed to find any evidence (Bange, Brown *et al*. 2001). Doubts over this assumption started to be raised however when an outbreak was reported at a cystic fibrosis clinic in Seattle in 2012, which involved five patients who all shared strains with an identical PFGE type (Aitken, Limaye *et al*. 2012).

In this study, 170 *M. abscessus* samples were sequenced from 31 patients at Papworth Adult Cystic Fibrosis Centre over four years. Prior to the initiation of this study only a single *M. abscessus* genome had been described, so a primary aim was to gain an overview of the population structure and diversity across the species. A secondary aim was to relate this population structure to the patients, and understand how their infections were acquired.

5.2. Methods

Since 2007, samples were collected from patients positive for *M. abscessus* and stored as frozen aliquots from MGIT samples. Following initial culture on solid media, sweeps of mycobacterial colonies were taken and sub-cultured (to remove contamination while maintaining genetic diversity). 170 pure cultures of *M. abscessus* were obtained from 31 patients (22 CF, 4 non-CF and 5 unknown), and DNA extractions were carried out. The DNA was subjected to 75bp paired-end sequencing on the Illumina HiSeq platform. Antibiotic sensitivity testing was performed on isolates before or shortly after starting anti-mycobacterial chemotherapy by serial broth microdilution and plates were read at 5 days.

For a species-wide analysis, the raw reads were mapped and SNPs called against the *M. abscessus* reference genome as described in Methods 8.2 and 8.3. In addition publically available data were analysed for isolates collected from Malaysia (M93 (Choo, Wong *et al*. 2012), M94 (Choo, Wong *et al*. 2012), M152 (Ngeow, Wong *et al*. 2012), M115 (Ngeow, Wong *et al*. 2012), M154 (Choo, Wong *et al*. 2012), M139 (Ngeow, Wee *et al*. 2012)), France (CCUG48898 (Tettelin, Sampaio *et al*. 2012), CIP108541 (Choi, Cho *et al*. 2012)), Birmingham (47J26 (Chan, Halachev *et al*. 2012)) and Brazil (GO-06 (Raiol, Ribeiro *et al*. 2012)). A high level of divergence was observed, so a separate analysis for each sub-species was carried out by generating two extra references to represent the *M. a. massiliense* and *M. a. bolletii* subspecies. These were produced by *de novo* assembly (using Velvet- see Methods 8.4) of the sequencing reads from appropriate samples, chosen on the basis of high coverage and central phylogenetic position. Mapping and tree construction as described in Methods 8.6 were then carried out separately for each of the subspecies. An in-house program (Croucher, Harris *et al*. 2011), was used to detect recombination and remove it from the dataset.

Bayesian inference was implemented in BEAST, a program used for Bayesian Markov chain Monte Carlo analysis of genetic sequences (Drummond and Rambaut 2007). BEAST was run on the dataset using both log-normal relaxed and strict clock models, for 100,000,000 states, excluding a 10% burn-in as described in Methods 8.7. The results quoted here were from runs with the best effective sample size (ESS), and

a log-normal relaxed clock. As substitution rate estimates in BEAST are dependent on a molecular clock signal being present, Path-O-Gen was used to informally assess the clock-likeness of the dataset using linear regression (See Methods 8.7). Monophyly was assessed by repeating the BEAST analyses above and reporting the "monophly statistic" for certain clades. This was carried out in triplicate, and the number of times a clade was reported as being as monophyletic was counted, with the burn-in excluded.

Epidemiological analysis was carried out by Dorothy Grogono and colleagues at Papworth Hospital. For this, all visits to the hospital were extracted from the patients' notes and cross-referenced. Public Health England (I Roddick and T Inns) carried out statistical analyses comparing clustered and non-clustered isolates and their association with specific ward visits.

5.3. Results

5.3.1. Species-wide overview

A maximum likelihood tree of all the samples in the dataset is shown in Figure 24. This revealed the presence of three major clades that represent the three main subspecies most often described as *M. abscessus* subsp. *abscessus*, *M. abscessus* subsp. *bolletii* and *M. abscessus* subsp. *massiliense*. Although previously proposed to be separate species (Blauwendraat, Dixon *et al*. 2012), the whole genome phylogeny shows that the average nucleotide identity (ANI) between representatives is 99.1%; well above the 94% ANI previously associated with the species boundary (Konstantinidis and Tiedje 2007), suggesting they most likely represent subspecies. Of the 31 patients, 13 were infected with *M. a. abscessus*, 15 with *M. a. massiliense* and two with *M. a. bolletii*. One patient was co-infected with both *M. a. abscessus* and *M. a. massiliense* and was excluded from further analysis. The phylogeny was also supplemented with publically available CF and non-CF isolates from the UK, Brazil and Malaysia, which were distributed broadly across the tree, indicating that the Papworth sample collection is likely to be representative of the species.

5.3.2. Recombination

Manual observation of the whole genome alignment revealed sequences likely due to recombination between and within the sub-species. This could not be removed in an automated fashion due to the scale of the dataset, so variation due to recombination was maintained in Figure 24. For individual clusters recombination was detected and removed before analysis using a phylogenetic approach based on SNP density (Croucher, Harris *et al*. 2011). No recombination within patients was detected.

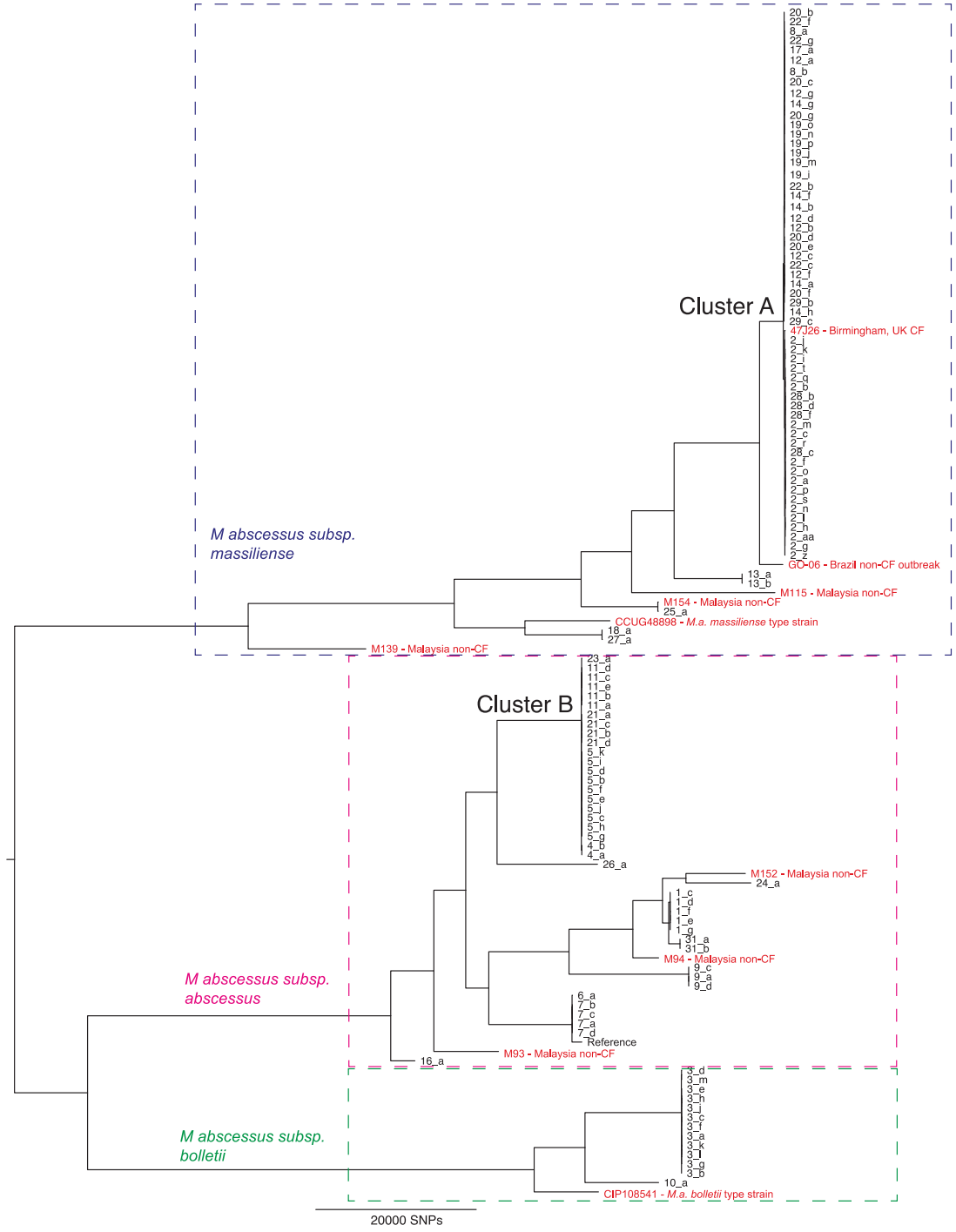


Figure 24 – Maximum likelihood phylogeny of all isolates in the collection (black) and publically available sequences (red). The three subspecies are indicated by the dashed boxes. The tree was built based on high quality SNPs called against the *M. a. abscessus* reference genome (labeled "Reference" in the tree). Possible transmission clusters A and B are indicated and shown in greater detail in Figure 25.

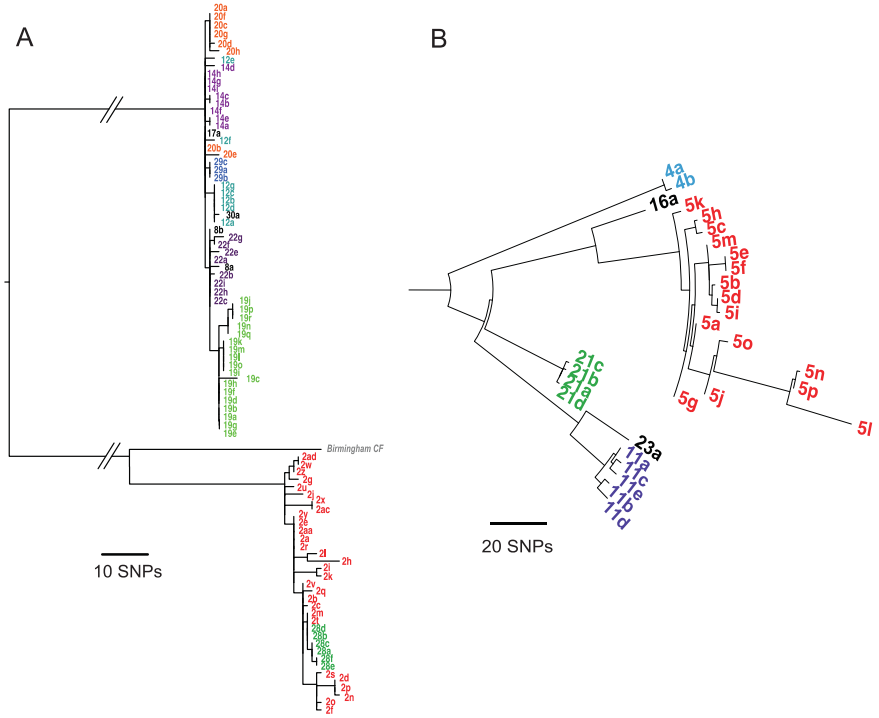


Figure 25: Maximum likelihood trees of cluster A and B as indicated in Figure 24. Trees were built based on SNPs called via mapping to a representative from each cluster. The bootstrap support for the sub-clusters in cluster A were 100%. Dashed lines represent branches shortened for illustration purposes.

5.3.3. Patient-level diversity

A detailed analysis was carried out for each subspecies, which showed that all patients had a clonal *M. abscessus* infection, as all their isolates were tightly clustered together on the tree. There were two clusters of multiple patient isolates (Cluster A and B: Figure 24) found in two of the subspecies, *M. a. massiliense* and *M. a. abscessus* respectively. Cluster A was comprised of two distinct sub-clusters separated by 185 SNPs, the first of which was composed of highly-related or identical isolates from nine different patients (Figure 25). The second of these was composed of two patients' isolates, with the diversity of one nested within the other: a pattern consistent with a directional patient-patient transmission event (Vandamme and Pybus 2013). The two SNPs (codon 42 in MAB_0477 and codon 233 in MAB_3748) conferring the monophyletic positioning of patient 28 within patient 2's isolates were found to be unique to these patients and of high quality. Furthermore the support for this reciprocal monophyletic topology was further supported by Bayesian analysis, discussed below in section 5.3.5.

In order to put the diversity observed within these clusters into context, the dataset was interrogated further by plotting the pairwise SNP distances (also known as Hamming distance (Pilcher, Wong *et al*. 2008)). This showed that whilst within patients most SNP distances are under 20 (red- Figure 26b), distances between patients mostly exceed 1,000 SNPs (not shown), which is consistent with independent acquisition from the environment. However, for the patients within *M. a. massiliense* cluster A (blue- Figure 26b), the inter-patient distances were very small, often the same or less than the number of SNPs seen within patients. There was also a second "mode" spanning 50 to 200 SNPs which represented the distance between the sub-clusters in *M. a. massiliense* cluster A in addition to the diversity within *M. a. abscessus* cluster B. Thus it appears that cluster B is "looser", with its inter-patient relationships more distant than observed for cluster A. However this diversity is still distinct from the diversity observed across the rest of the dataset as a whole (Figure 26).

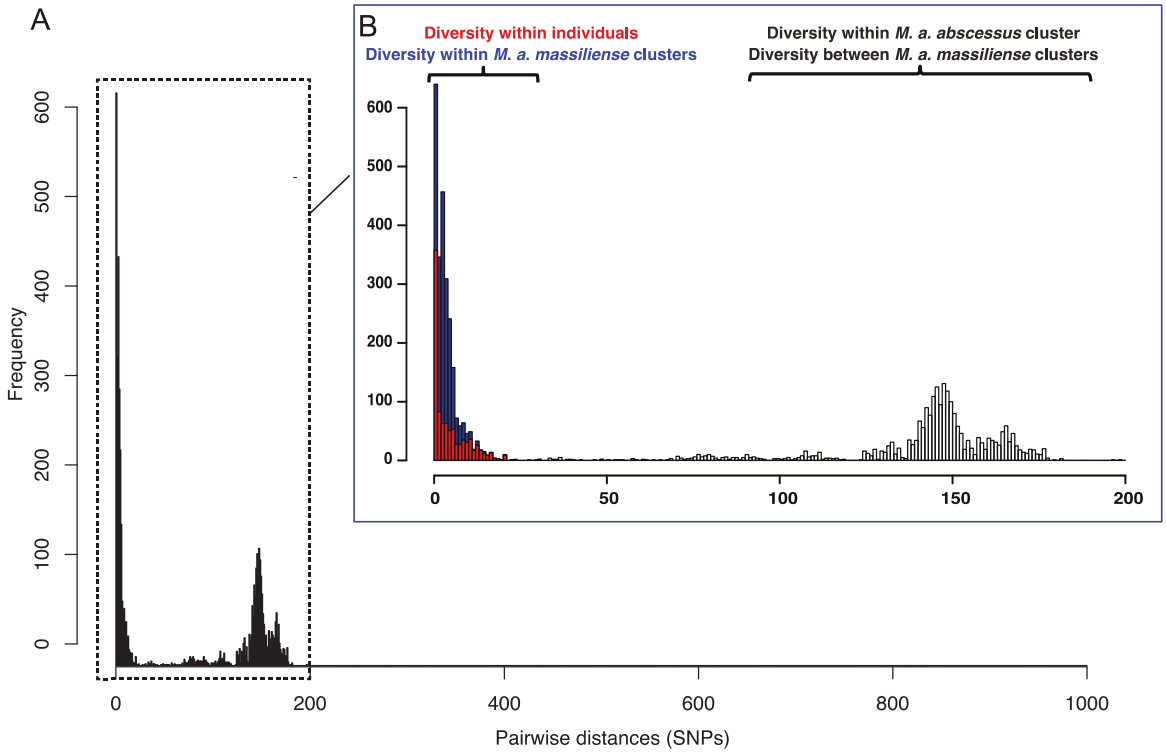


Figure 26: Histogram of pairwise SNP differences. A: all SNP differences under 1000. B: all SNP differences under 200 with subsets of data labeled. B shows two distinct "modes" existing in the data and A demonstrates how distinct these are from the rest of the dataset (all > 1000 SNPs).

5.3.4. Epidemiological investigation

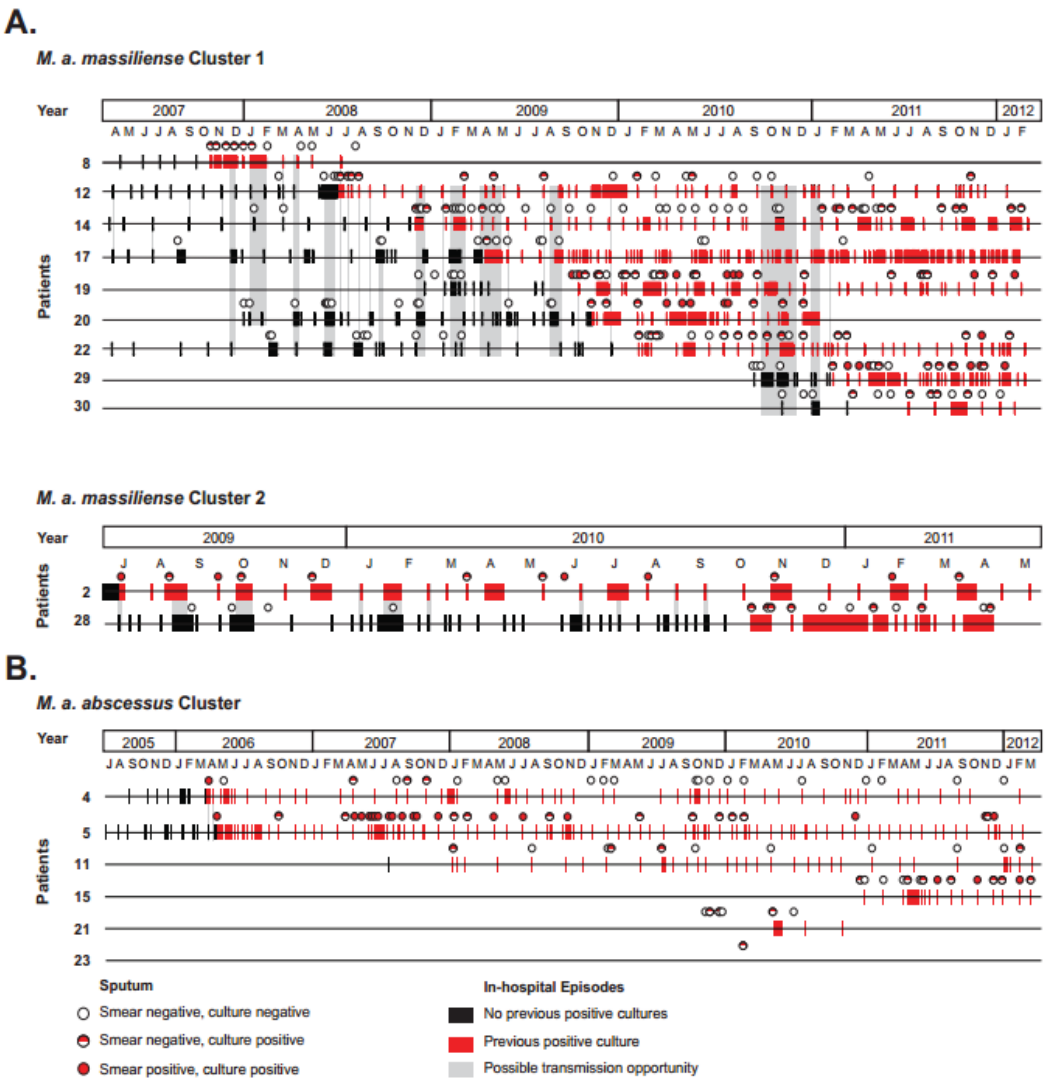


Figure 27 – Opportunities for transmission between patients. The timelines of individual patients within the *M. a. massiliense* sub-clusters 1 and 2 (A) and *M. a. abscessus* cluster B (B) are shown with short vertical lines denoting hospital visits to Papworth Hospital, and circles denoting sputum samples (culture negative white; smear negative culture positive half-red; smear positive red). Timelines become red following a positive sputum sample. Potential opportunities for transmission between patients (negative patient in hospital at the same time as a positive patient) are highlighted by grey vertical bars.

The tight relationships between different patients' isolates raised the possibility that either a point-source outbreak or patient-patient transmission was occurring. To investigate the former, Papworth hospital initiated extensive environmental sampling in June 2010 when the genomic analysis had revealed the possibility of an outbreak. The hospital water supply, which is chlorinated on site, was extensively sampled and

repeatedly found to be culture and PCR-negative for Mycobacteria. Showerheads, dishwashers and bronchoscopes were also shown to be free from NTM. In addition there was no geographical association between the patients' home water supply and genetic clustering.

To investigate the possibility of patient-patient transmission, Papworth hospital collected information from the patients' notes to determine when and where they attended the hospital. They could not find occasions, including social links, outside the hospital where direct patient-to-patient transmission might have occurred, but did, however, identify clear opportunities for transmission within the centre for all patients from the two *M. a. massiliense* sub-clusters (Figure 27). Except for the presumed index cases (patients 8 and 2 for sub-cluster 1 and 2 respectively), all previously uninfected patients were present at the centre at the same time as an infected individual on multiple occasions. In contrast, patients infected with the *M. a. abscessus* cluster B isolates had no clear opportunities for transmission within or outside the hospital, with the only overlap identified between patient 4 and 5 whose isolates are not monophyletic in the phylogeny (Figure 25).

This epidemiological analysis was extended further in order to determine whether patients with isolates belonging to the *M. a. massiliense* cluster had more opportunities for transmission than patients with non-clustered isolates. Individuals within *M. a. massiliense* sub-cluster 1 (n = 9) were compared to patients with unclustered *M. abscessus* isolates (n = 15). The incubation period of *M. abscessus* infections is unknown so it was assumed that patients might acquire infection any time during a 12-month period before their first positive sample and might transmit infection at any time from this point onwards. For every 100 days during periods of potential acquisition, clustered cases were significantly more exposed to hospital than unclustered cases (mean 10.8 days vs. 4.1 days; $p = 0.0126$), had greater exposure to the CF inpatient ward (mean 5.7 days vs 1.5 days; $p = 0.0133$) and were more likely to be in hospital at the same time as potentially infected individuals (4.18 days vs. 0.63 days; $p = 0.0053$).

5.3.5. Bayesian dating of possible patient transmission events

Next, dating techniques were used in order to estimated the ages of the two clusters, and whether they were consistent with the opportunities for transmission identified in the epidemiological analysis. Using BEAST (Drummond and Rambaut 2007) with a constant population size and lognormal relaxed clock, an estimated substitution rate of 1.8 (0.3-3.3 95% highest posterior density) was estimated for *M. a. abscessus* cluster B and 0.47 (0.2-0.8 95% HPD) SNPs per genome per year for *M. a. massiliense* cluster A. Estimates were also made for the ages of the most recent common ancestors (MRCAs) of linked patients found adjacent on the tree. In the case of both *M. a. massiliense* clusters, the estimates for the age of the inter-patient MRCAs overlapped with the opportunities for hospital transmission (Figure 28) For *M. a. abscessus* cluster B this was not the case as its inter-patient MRCA of patient 4 and 5 dated to several decades prior to either of the patients becoming positive (Figure 28).

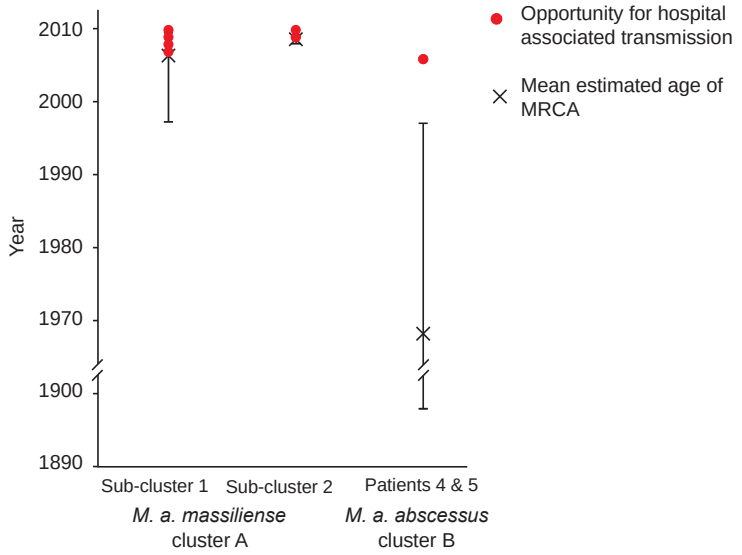


Figure 28 – Agreement between opportunities for patient-patient transmission and the estimated age of the corresponding most recent common ancestor (MRCA). Estimates from BEAST (Drummond and Rambaut 2007), of when the MRCA existed for isolates from different patients within *M. a. massiliense* clusters 1 and 2 and from the two patients within the *M. a. abscessus* cluster who had transmission opportunities

The substitution rates and the age of MRCAs presented here depend on the presence of a molecular clock in the dataset. In general a positive significant correlation between patient diversity and time was observed (see section 6.3.1), suggesting the

presence of a molecular clock. For individual clusters, Path-O-gen (Rambaut 2007) was used to plot root-tip distances against time to assess the presence of a strict clock-like signal. For the *M. a massiliense* sub-clusters, this test performed poorly, with a very weak positive signal in *M. a. massiliense* sub-cluster 1 (Figure 29a) and a complete absence in *M. a massiliense* sub-cluster 2 (Figure 29b). As *M. a. abscessus* cluster B has much deeper relationships between the patients, each patient (with more than 2 isolates) was plotted separately. In this case a positive molecular clock signal could be detected (Figure 30). The lack of a molecular clock observed in the *M. a. massiliense* data weakens the conclusions that can be drawn from the BEAST analysis. However it should be noted that the root-to-tip analysis is not a formal statistical test of a molecular clock, due to non-independence, so has its own limitations.

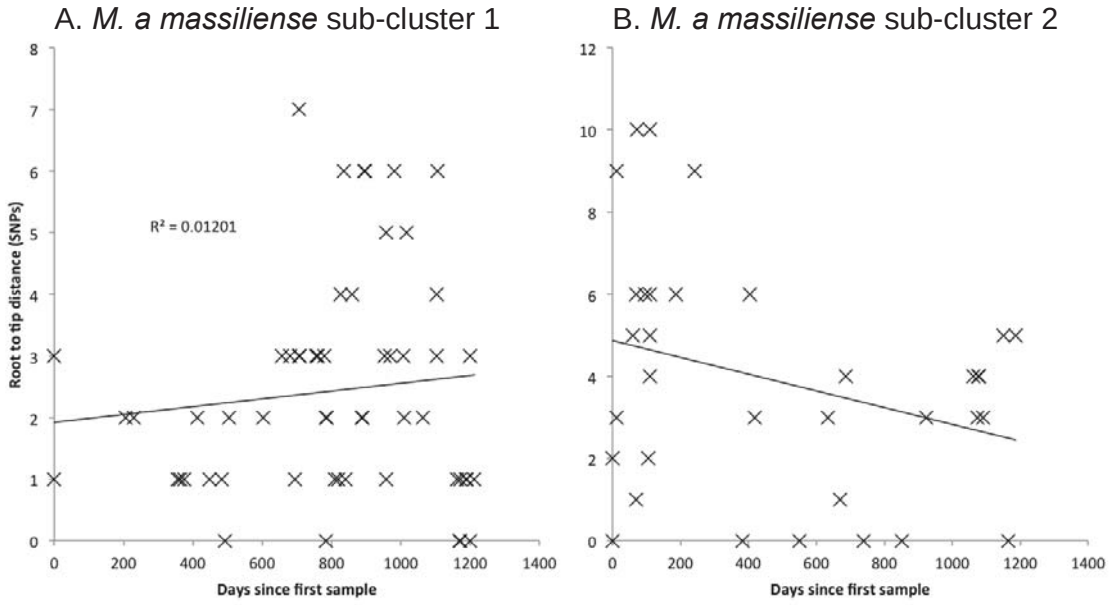


Figure 29 - Root to tip distances for *M. a. massiliense* sub-clusters. Rooted using Path-O-Gen (Rambaut 2007). No correlation coefficient is stated for sub-cluster 2 as the correlation is negative and shows a complete lack of evidence for a molecular clock.

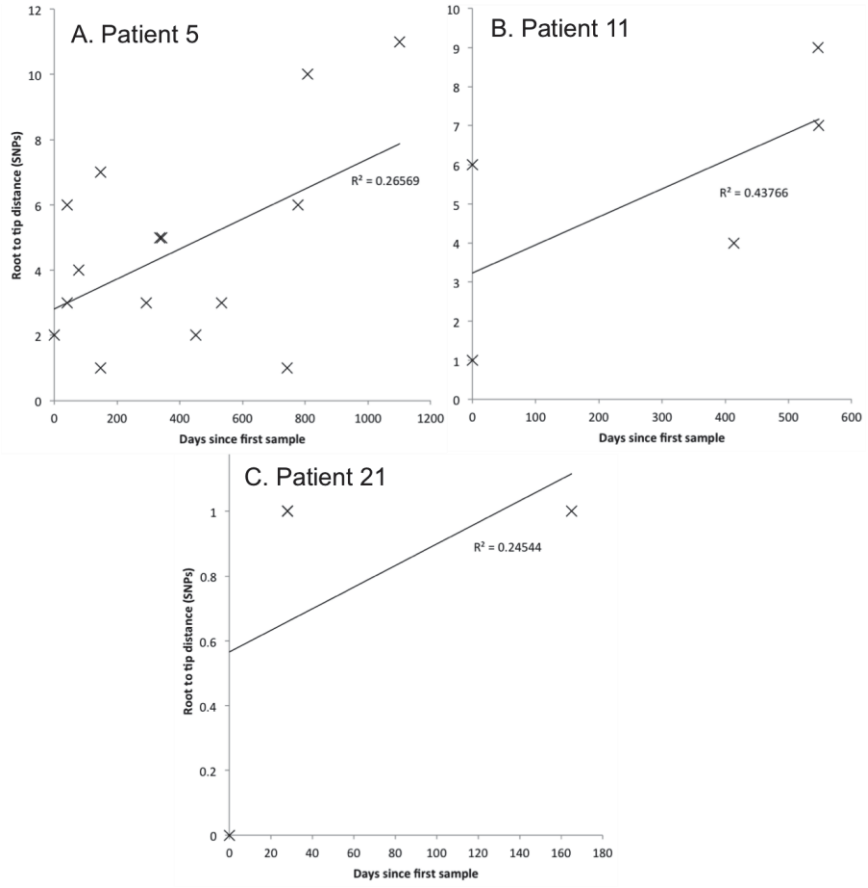


Figure 30 - Root to tip distances for individual patients within *M. a. abscessus* cluster. Rooted using nearest neighbor.

The reciprocal monophyly observed for patients 2 and 28 in *M. a. massiliense* sub-cluster 2 was tested further using BEAST. The "monophyly statistic", which determines for each state whether a certain group of taxa are monophyletic or not, was reported every 1000 states out of a total chain length of 100,000,000. This was carried out for isolates belonging to patient 28, isolates belonging to patient 2 and both patients combined. This confirmed that by far the most common scenario sampled is where patient 28 and patient 2 combined are monophyletic (Figure 31), and that patient 2 is never monophyletic, due to the nesting of patient 28 within it. This provides confidence to the observation of a nested topology involving these patients, a pattern strongly indicative of patient transmission.

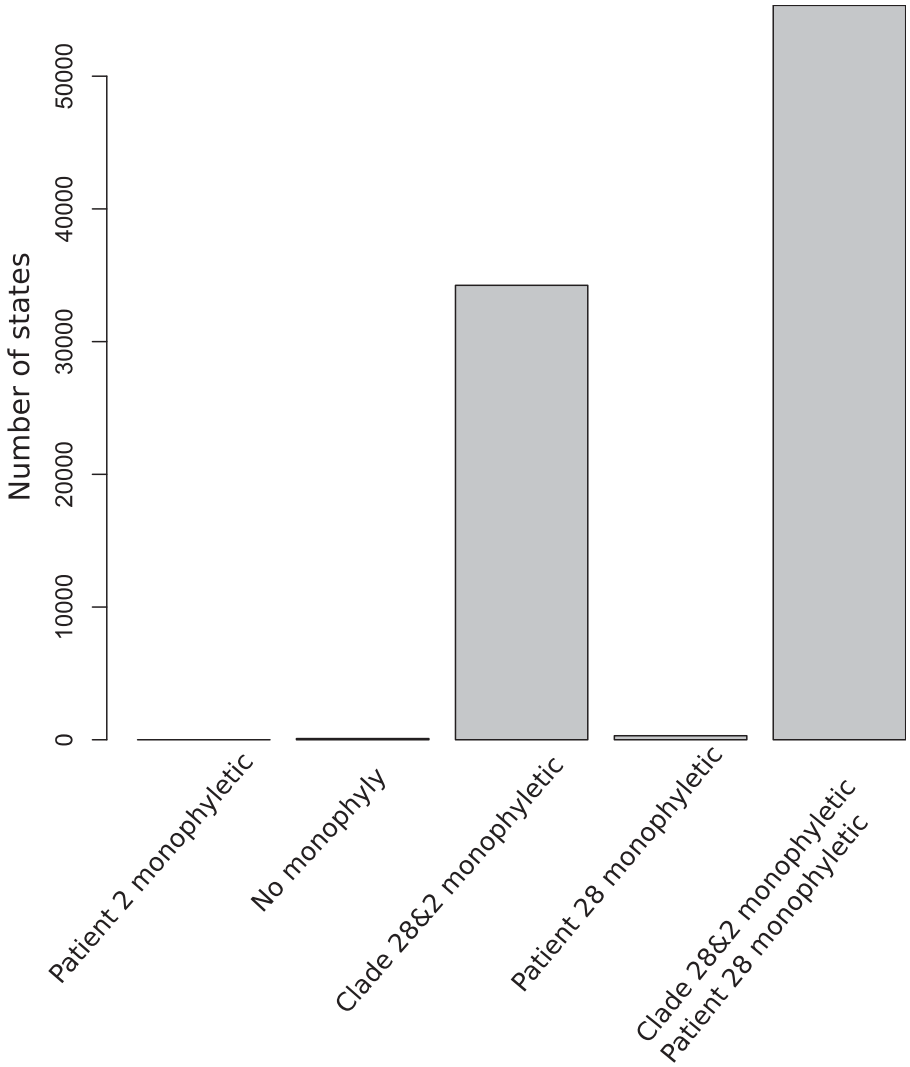


Figure 31 - Number of tree topologies where patients' isolates are monophyletic within *M. a. massiliense* sub-cluster 2. BEAST 1.7.5 was run on *M. a. massiliense* cluster A dataset described in the text with a chain length of 100 million states. The monophyly statistic was sampled every 1000 states, and a 10% burn-in was discarded. Patient 2, patient 28 and a clade comprising of both patient 2 28's isolates were tested. The number of times each scenario was counted is shown. Two additional runs on the same dataset showed highly similar results and are included in Appendix 9.5.

5.3.6. Antibiotic resistance patterns of *M. a. massiliense* cluster A

Amikacin and clarithromycin are commonly used to treat *M. abscessus* infections, and the molecular bases of acquired resistance to them are well characterised. Resistance to amikacin, an aminoglycoside, most often occurs through mutations in the 16s ribosomal RNA (Prammananan, Sander *et al*. 1998). Inducible clarithromycin (macrolide) resistance occurs through the up-regulation of the *erm(41)* gene (Nash, Brown-Elliott *et al*. 2009). Previously it was found that in *M. a. massiliense* strains, part of the *erm(41)* gene is deleted, rendering it inactive (Nash, Brown-Elliott *et al*.

2009). Instead constitutive resistance can evolve in *M. a. massiliense* through mutations in the 23s ribosomal RNA (Nash, Brown-Elliott *et al*. 2009). It was found that both *M. massiliense* cluster A sub-clusters were phenotypically and genotypically highly resistant to clarithromycin, but through independent mutations (Figure 32). This kind of acquired resistance is only thought to occur in patients upon exposure to macrolides (Wallace, Meier *et al*. 1996), and this assumption is supported by preliminary analysis of a global collection of *M. abscessus* isolates (manuscript in preparation), where all resistance conferring mutations were observed to occur at or close to the tips of the tree, and were not sustained in the population (Appendix 9.5, Figure 50). As three of the patients in sub-cluster 1 had not been exposed, this is perhaps evidence of human related transmission, assuming that resistance was acquired by a patient prior to sampling. Only sub cluster 1 was resistant to amikacin. However, the records of patient 8 (the presumed index case based on date of acquisition) revealed that they had acquired resistance during treatment before sampling was initiated. This suggests that resistance was acquired by this patient and then inherited by other patients through transmission.

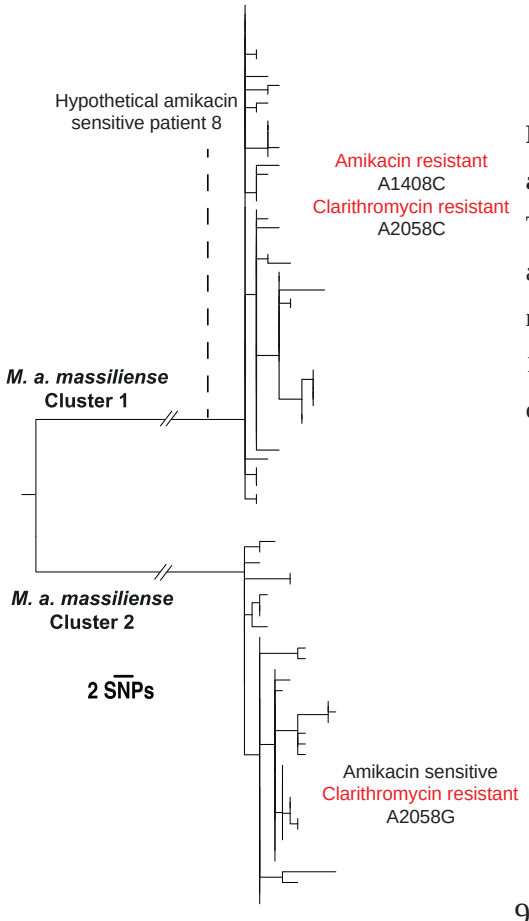


Figure 32 - Antibiotic resistance phenotype and genotypes of *M. a. massiliense* cluster 1. The dotted line indicates the hypothetical amikacin sensitive ancestor of cluster 1. When resistant, the causal mutation is shown in the 16S and 23S genes for amikacin and clarithromycin resistance respectively.

5.3.7. Evidence for dominant circulating clones

In contrast to the sub-clusters within cluster A, the possibility of patient-patient transmission in cluster B was unsupported by both its genetic diversity and epidemiology. However, the level of diversity observed (50-200 SNPs) is still lower than and distinct from the rest of the dataset, suggesting that this cluster is the result of a different underlying process. It's possible that this cluster represents a dominant circulating clone that may be the result of past transmission events amongst unsampled patients or a clone that is more prevalent in the general and/or cystic fibrosis population than would be expected by chance. In addition, the *M. a. massiliense* cluster A might also represent a dominant circulating clone, as the distance between the sub-clusters also falls within the 50-200 SNPs range. Of the published genomes, an isolate from a cystic fibrosis patient in Birmingham sits within cluster A between the two sub-clusters, and an isolate from a large outbreak in Brazil also appears to be ancestral but closely related to the cluster (Figure 24). Both of these sub-clusters are MLST type 23 and are found 40 times in the *M. abscessus* MLST database (total size 284, April 2014) maintained by the Pasteur Institute. The *M. a. abscessus* cluster B is MLST type 26 and is only found once on the database, from an isolate in France. Although the database is by no means extensive, it demonstrates that these putative dominant circulating clones have a further reach than Papworth or the UK, but that further sampling and characterization is required to determine their spread and nature.

5.4. Discussion

This analysis represents the first application of high-throughput genomics to *M. abscessus*, an important cystic fibrosis pathogen. The most significant finding of this work was that there was strong evidence for transmission between patients within a single cystic fibrosis clinic. NTMs are commonly cited as non-transmissible so cases aren't typically scrutinised for the possibility of cross-infection between patients. Therefore this work has an important clinical impact, and raises the possibility that other under-studied cystic fibrosis pathogens may also be able to transmit.

Identical or near-identical isolates of *M. abscessus* were observed in 11 patients in two sub-clusters, which was in contrast to the large genetic distances between isolates in the other patients. This could be due to either a point environmental source, or transmission between patients. Mycobacteria are notoriously difficult to culture from the environment (Falkinham 2002), and sampling occurred prospectively after the putative outbreak had been identified meaning that an environmental source cannot be completely ruled out. However there are several lines of evidence that suggest that a point source is unlikely. Firstly the point source would need to be maintained over a period of four years, and all of the patients would have needed to be exposed to this single source. The epidemiological analysis concluded that none of the clustered patients were associated with a single room or type of treatment. Secondly, the phylogenetic topology of one of the sub-clusters found in *M. a. massiliense* cluster A is highly consistent with directional transmission; where the diversity of the recipient (patient 28) is found to be nested within the diversity of the source case (patient 2), a pattern often observed for HIV transmission (Scaduto, Brown *et al*. 2010). This kind of pattern would be unlikely in a point-source scenario where a star shape, rather than a chain-like phylogeny would be expected (Ypma, Donker *et al*. 2013). Thirdly, the putative transmission clusters had acquired resistance, something that wouldn't be expected for an environmental bacterium, suggesting that there is at least some human element of the transmission chain. Therefore the most parsimonious explanation for these patterns is that transmission has occurred between these patients. This is counter to a previous study that found a complete absence of *M. abscessus* patient-patient transmission (Bange, Brown *et al*. 2001). However, the study was limited to a small number of patients (n=5) over a two year period, whereas this study encompasses a

much larger cohort (n=31) collected over four years, providing a greater opportunity to detect transmission.

Many of the patients involved in these transmission clusters were found to attend the hospital on the same day, perhaps providing an opportunity for cross-infection. However direct transmission between the patients is unlikely as strict infection control procedures are already in place at Papworth, preventing patients waiting in the same area or coming in direct contact within the hospital. Instead, indirect transmission is more likely, although in the absence of environmental evidence a specific mechanism can only be speculated upon. The ability of *M. abscessus* to withstand desiccation and other physical stresses and its resistance to many disinfectants (Wallace, Brown *et al*. 1998) may allow transmission *via* fomite contamination. Aerosol generation during physiotherapy and/or lung function testing could lead to the production and inhalation of airborne water droplets, from which NTM have been cultured in the environment (Wendt, George *et al*. 1980). Alternatively, it is possible that non-CF patients could be asymptomatic carriers, although there is currently no evidence to support this. It is also worth noting that transmission sometimes occurred between patients with persistently smear-negative (but culture positive) sputum, suggesting that the infectious dose may be low.

In addition to the patterns of diversity consistent with recent patient transmission, another distinct mode was detected that may represent the diversity seen within a dominant circulating clone. The existence of such clones is a well-recognised phenomenon of *P. aeruginosa*, where "epidemic" clones are found more often in cystic fibrosis patients than the environment and are thought to have a greater propensity to transmit (Wiehlmann, Wagner *et al*. 2007). Although there is still a possibility that *M. a. abscessus* or *M. a. bolletii* may be able to transmit, it is interesting that the only evidence of patient transmission was found within the *M. a. massiliense* dominant circulating clone (cluster A), raising the possibility that it may be particularly adapted for human spread. Subsequent to the work described here, several new whole genome analyses have been carried out on collections from outside the UK that strengthen the evidence for a dominant circulating clone. One study sequenced two strains from a large outbreak of *M. abscessus* of over 2,000 skin infections in Brazil, which are all assumed to be due to one PFGE-defined clone

(Leao, Viana-Niero *et al*. 2010). When comparing the Brazilian strains against the Papworth isolates presented here, they found they cluster closely with the *M. a. massiliense* cluster A (Davidson, Hasan *et al*. 2013). Surprisingly, although part of the same clone, they were relatively distant from the previously sequenced Brazillian strain GO-06, which suggests that the outbreak is likely to be polyclonal. One of these Brazilian strains (CRM-0019) has been shown to have a higher level of virulence in macrophages, than the type strain (Shang, Gibbs *et al*. 2011), which supports the idea that *M. a. massiliense* putative dominant circulating clone may be particularly adapted for human spread, although not necessarily through respiratory routes. Another recent study applying whole genome sequencing to the previously described outbreak in Seattle (Aitken, Limaye *et al*. 2012), also found that their strains clustered closely with, but were not identical to, the *M. a. massiliense* putative dominant circulating clone (Tettelin H, Davidson R.M *et al*. 2014). This again strengthens the hypothesis that these dominant circulating clones exist, and may be particularly clinically relevant. A larger and broader study will be required to determine their reach and nature.

The finding of frequent transmission amongst patients with cystic fibrosis raises several important questions about current infection control measures used in cystic fibrosis centres, as noted by commentators (Elborn 2013, O'Sullivan and Sassetti 2013). In response to these findings Papworth hospital has implemented new infection control measures including: continuous sputum screening for NTM in all patients; outpatient segregation of infected patients within a dedicated outpatient clinic with single use rooms; and use of negative pressure rooms for inpatient care. Follow up sequencing will be required to determine whether this has prevented further transmission.

6. Within-patient evolution of *Mycobacterium abscessus*

This work forms the basis of a manuscript in preparation:

J. M. Bryant, D. M. Grogono, U. Hill, C. S. Haworth, J. Foweraker, J. Parkhill and R. A. Floto . Tracking genetic diversity and evolution of *Mycobacterium abscessus* during chronic infection using whole genome sequencing.

Statement of contribution

I carried out all bioinformatic analyses. DMG carried out collection of clinical data, antibiotic susceptibility testing and colony phenotyping. Study was supervised by JP and RAP. All authors contributed to interpretation of the data.

6.1. Introduction

Infections of cystic fibrosis patients are often extremely chronic and difficult to treat, sometimes leaving patients permanently colonised for the rest of their lives. These chronic respiratory infections present an evolutionary scenario where the length of colonisation allows the development of a high level of within-patient diversity and the opportunity to adapt to the cystic fibrosis lung niche. This has been well documented for *P. aeruginosa* and *Burkholderia dolosa*, where processes such as colony morphology switches (Govan and Deretic 1996), hypermutation (Oliver, Canton *et al*. 2000), selective sweeps (Lieberman, Flett *et al*. 2014) and parallel genetic mutations (Lieberman, Michel *et al*. 2011) between patients have been documented. However, how the progression of this diversity over time correlates with clinical phenotype or treatment is yet to be fully quantified and has only been investigated in a small number of patients (Workentine, Sibley *et al*. 2013). In particular, the within-patient diversity of *M. abscessus* is largely unexplored.

Like most cystic fibrosis pathogens, *M. abscessus* is thought to be primarily an environmental bacterium, which means upon entering the human host, genetic and transcriptional changes may be required to allow the bacteria to thrive in the new niche. The most apparent evidence for this process is the spontaneous transformation from a smooth to rough colony morphology, with the latter considered more virulent and able to thrive within the human host. Other than one small scale study (Kreutzfeldt, McAdam *et al*. 2013), the within-host variation of *M. abscessus* infections has been unexplored. Considering that drug resistance is often acquired within patients (Nessar, Cambau *et al*. 2012), and the obscure and irreproducible results of drug susceptibility testing for some antibiotics (Broda, Jebbari *et al*. 2013), it is important that this is fully understood.

Most bacterial whole genome studies to date have been based on colony-purified samples; so have focused on calling consensus SNPs as all variants should be fixed in the sample. This has meant there is little opportunity to investigate within-patient diversity through the detection of unfixed minority variants. This study aimed to capture the within-patient diversity of *M. abscessus*, utilising 151 whole genome sequences from 21 cystic fibrosis patients, described in a previously published data-

set (Chapter 5). To capture as much genetic variation as possible, the samples were not colony-purified, allowing the detection of minority variants and an estimation of the level of diversity present within the patient at each time point. This fine-scale genetic information was correlated with clinical phenotype and treatment, which provided insights into the diversity of *M. abscessus* and chronic bacterial infections in general.

6.2. Methods

6.2.1. Samples

For this study, a subsample of the sequencing data from a previously described dataset was used (Chapter 5). This was limited to patients diagnosed with cystic fibrosis only (n=22, 155 samples). One patient's data (4 samples) were excluded as they were found to have a mixed infection comprising of *M. a. massiliense* and *M. a. abscessus*. For the minority variant analysis, only patients with five or more samples collected at least seven days apart and no evidence of contamination were used resulting in a final dataset of 112 isolates from 10 patients (see appendix 9.5)

6.2.2. Detection of minority variants

Sequencing was carried out on DNA extracted from sweeps of *M. abscessus* maintaining a sample of the within-patient diversity. In order to detect minority variants, where all reads do not agree on a consensus base, additional steps to those used for normal SNP calling are required. As a first step stringent mapping was applied, where in addition to the default SMALT parameters, a minimum nucleotide identity of 0.98 was used, which avoided the mapping of reads with more than one miss-match which could be considered poor quality. The resultant variant data was then filtered for high quality minority variants and using parameters as described in Methods 8.9. Numbers of minority variants were corrected for coverage as described in Methods 8.9.

6.2.3. Phenotyping of colony variants and antibiotic susceptibility

Phenotyping was carried out by Dorothy Grogono at Papworth hospital. Isolates were streaked onto blood agar to examine colony morphology, and were recorded as either

Smooth [S], Rough [R], Rough/ Smooth [R/S] (where an isolate displayed colonies of both morphotypes), or Indeterminate [I].

Individual colonies were picked from solid media and grown up as pure cultures and antibiotic susceptibility testing was performed by serial broth microdilution (using standard CLSI methods (Institute 2011)). Plates were read at 3 - 5 days, and clarithromycin results were also read at 14 days to detect inducible macrolide resistance.

6.2.4. Detection of polymorphisms within genes associated with antibiotic resistance

A literature search was carried out in order to identify possible drug targets and genes associated with resistance to aminoglycosides, cephalosporins, quinolones, macrolides, tetracycline and carbapenems. In addition any genes annotated as beta-lactamases or penicillin binding proteins were also added to the list. The final list can be found in Appendix 9.5. All of the non-synonymous minority variants and consensus SNPs generated within patients (i.e. not shared between all samples in a patient) were crosschecked against this list.

6.2.5. Detection of polymorphisms associated with colony morphology switches

Several different approaches were taken to identify candidate polymorphisms. Firstly, all minority variants or consensus SNPs that differed between isolates of the same patient with a clear differentiation between the rough and smooth morphotypes were identified. Secondly, as heterogenous indels are difficult to call using standard methods, any homopolymers (>4 nt) occurring in the GPL loci were detected using a custom perl script. The length of the homopolymer found in each raw read was then extracted from the mapping data in order to identify minority variants conferring frameshift mutations.

6.3. Results

Overall, 107 consensus SNPs were detected within the patients (variants where over 80% of reads agree) using standard SNP calling methods. An additional 899 minority variants (where 4 or more high quality reads disagree with the consensus) were also detected, of which 630 were unique (only appear in one sample). This demonstrates that significant within-sample variation exists which can be detected using a high depth of coverage (average for this dataset 118 fold).

6.3.1. The number of minority variants can be used as a proxy for population size

When considering consensus SNPs only, the average pairwise diversity between isolates from the same patient was found to correlate positively with the length of their infection (Figure 33) This reflects an increase in diversity over time, which may represent the accumulation of variants being fixed in the population occurring at a rate of approximately 1 SNP per genome per year (derived from the slope of the linear regression model fitted). For minority variants, a positive trend was also observed, however this relationship was not significant (Figure 33).

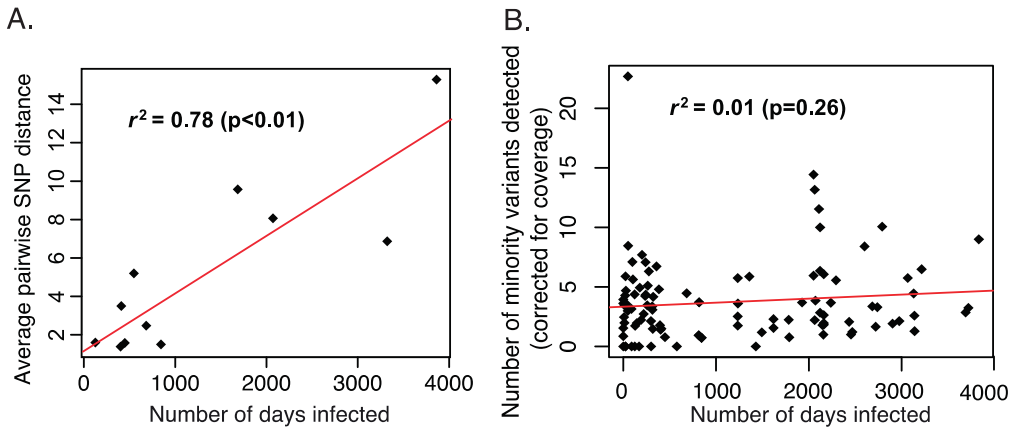


Figure 33 - Relationship between time and diversity. A) Total time since first positive *M. abscessus* sample vs. average pairwise SNP distance for each patient. Only patients with 5 or more sequenced samples available were included. B) Time vs. number of minority variants detected for each isolate. The number of minority variants detected was corrected for the depth of coverage. With the outlier removed (22 minority variants) the P value of the linear regression model reaches significance (P = 0.042).

This lack of significance could be the result of a complex array of factors affecting the bacterial population at any particular time point in the infection. These different factors are further explored in the rest of this chapter.

As evolutionary theory predicts that mutations will occur more frequently in larger populations, the observed changes in the number of minority variants over time could reflect changes in population size. In order to test this, the population size can be measured indirectly using an automated MGIT culture system (BACTEC™ MGIT™ 960), where the time taken to positivity has been shown to correlate negatively with bacterial load (Diacon, Maritz *et al*. 2012). Supporting this, the number of minority variants was found to correlative negatively with time to positivity for individual patients (Figure 34). However, isolates with a rough colony morphology type were excluded from this analysis, as they were observed to excessively clump in culture, which may have interfered with the automated detection system. In addition, when the data was combined any significant negative correlation was lost, which may be due to strain specific differences in their ability to grow in this culture system.

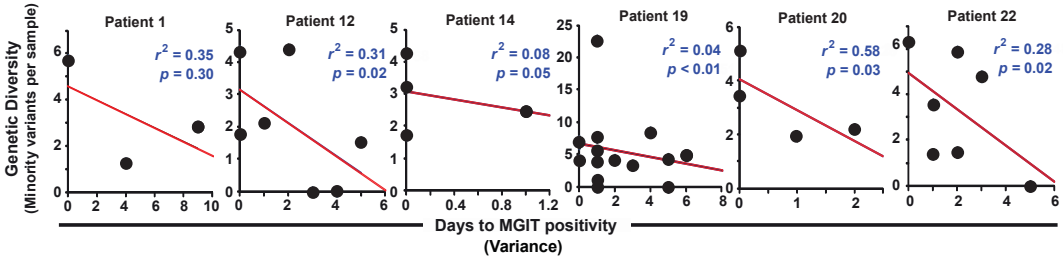


Figure 34 - Days to MGIT positivity vs. genetic diversity. The time taken for the Mycobacterial growth indicator tube to flag positive is known to correlate negatively with bacterial burden, as does the genetic diversity (number of minority variants – corrected) as shown here. Only patient's isolates with a smooth colony morphology were included.

6.3.2. Genetic diversity correlates with infection severity

For many patients, there were large fluctuations in the number of minority variants over the course of their infection. These were plotted against clinical markers of infection (C-reactive protein and Forced Expiratory Volume in 1 second) and antibiotic treatment to see if they indicted factors that might explain these changes in diversity and therefore population size (Figure 35).

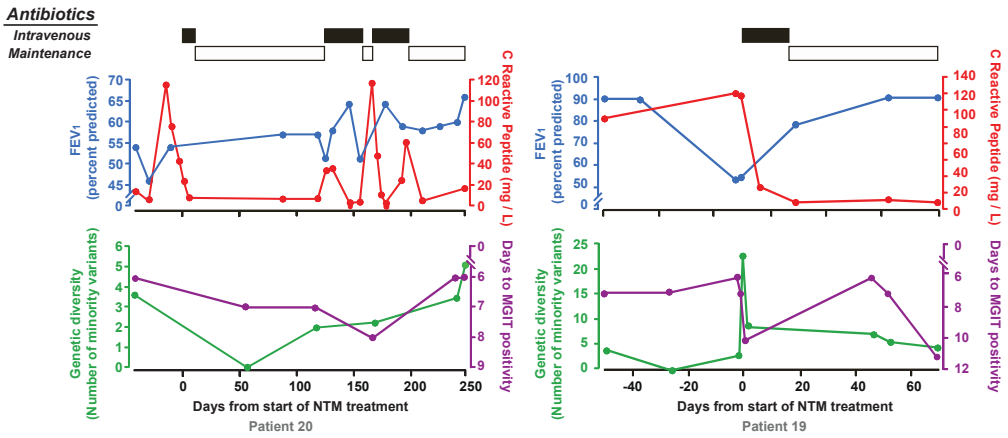


Figure 35 – Changes in genetic diversity and clinical phenotype in two patients over time. Time lines for two patients illustrating how changes in genetic diversity (tracked by the number of minority variants per isolate; green) relate to lung function (measured by FEV1; Blue), systemic inflammation (measured by C Reactive protein; red), bacterial burden (which is inversely related to days to MGIT positivity; purple) as well as maintenance (white bars) and intravenous (black bars) antibiotic therapy.

There were multiple occasions where spikes in the number of minority variants occurred at the time of an exacerbation of infection (substantiated by a rise in C-reactive peptide, fall in FEV1 or both) (Figure 35). Across all patients, CRP levels were found to correlate with the number of minority variants, presumably because both are influenced by changes in population size (Figure 36a). In addition to this correlation, it was found that successful antibiotic treatment (usually associated with administration of intravenous antibiotic therapy) was associated with decreases in the number of minority variants (Figure 36b), suggesting the occurrence of evolutionary bottlenecks. It is worth noting however that this reduction is often not maintained in subsequent samples, suggesting the population size can recover after initial exposure to the drug.

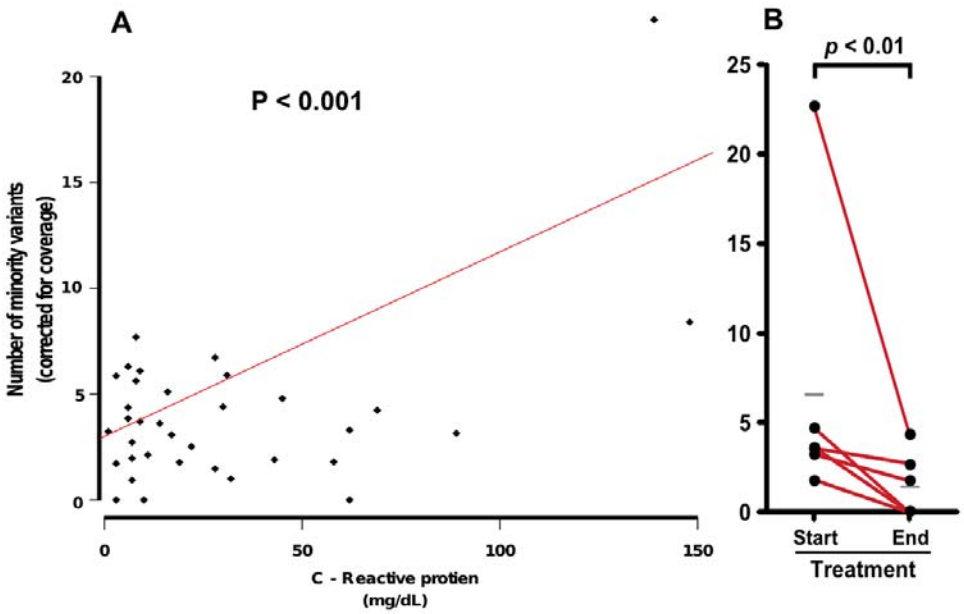


Figure 36 - Correlation between genetic diversity, CRP and treatment. A) Relationship between within sample genetic diversity (number of minority variants) and systemic inflammation of infected patients (monitored by CRP). Calculated using Poisson regression. The correlation was still significant (P < 0.001) with the outlier (MV count greater than 20) removed. B) Effect of successful antibiotic therapy (resulting in improved FEV1 and C Reactive Protein) on within sample genetic diversity. P value calculated using the paired Wilcoxon signed-rank test.

6.3.3. Population bottlenecks occur through patient-patient transmission

As described previously (Chapter 5), there was strong phylogenetic and epidemiological evidence supporting person-person transmission from patient 2 to 28. For patient 2, the number of minority variants increased over time, indicating an increase in population size. For patient 28 however, lower numbers of variants were detected, a pattern indicative of a population bottleneck that may have occurred during transmission (Figure 37).

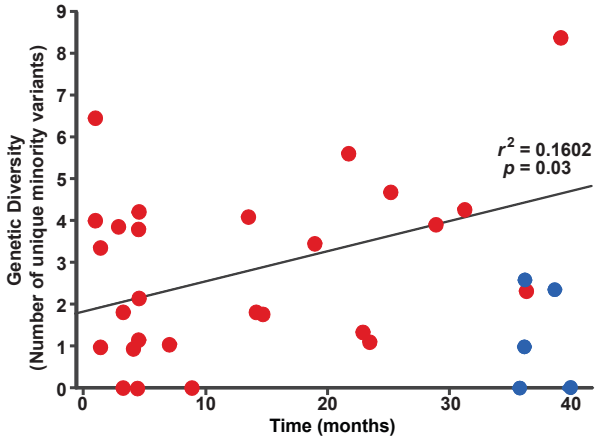


Figure 37 – Number of minority variants over time in patient 2 (red) and patient 28 (blue). Correlation coefficient and P value are only calculated based on patient 2's isolates.

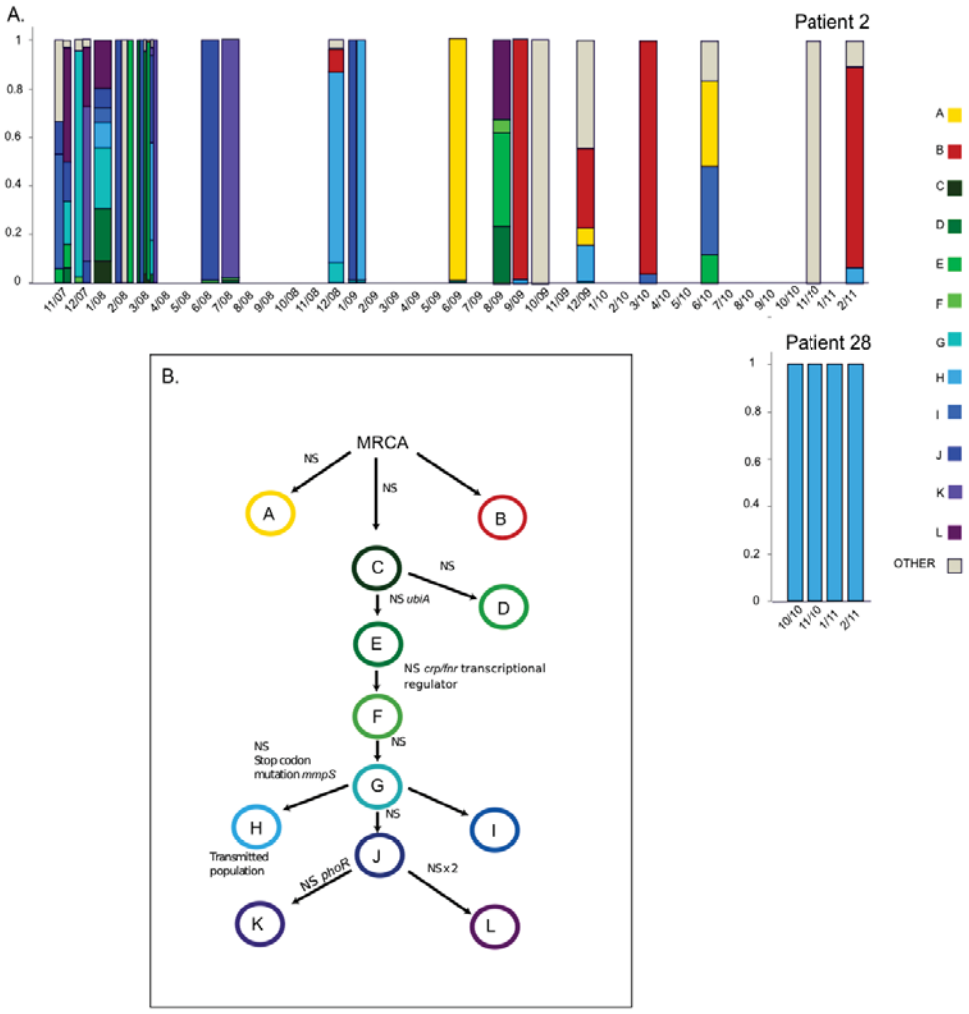


Figure 38 – Sub-clones identified in patient 2 and 28. A) Frequency of each subclone within each isolate analysed during infection of Patient 2 and transmission to Patient 28. This was calculated by determining the frequency of the "terminal" variant marking this subclone as shown in: B) Reconstruction of the lineages found in patient 2 (and transmitted to 28) demonstrates successive acquisition of non-synonymous polymorphisms (NS) by the most recent common ancestral clone (MRCA) in potential virulence genes (*ubiA*, MAB_0173; *crp/fnr*, MAB_0416c; *mmpS*, MAB_0477; *phoR*, MAB_0674).

Most minority variants detected within patients in this study were unique (only detected in one sample). However in patient 2, 17 variants were found to occur multiple times in different samples, and 14 of these co-occurred with one or more other variant at the same frequency (see Appendix 9.5), indicating linkage on the same genetic background. This suggests the presence of several sub-clones within the patient, which could perhaps be attributed to the length of the infection (exceeding 10 years). All of the variants were non-synonymous or intergenic and many were found in genes associated with virulence or drug resistance. Using the observed linkage

between these variants, hypothetical sub-lineages could be reconstructed (Figure 38b). The evolutionary succession of these lineages and the variants associated with them could be inferred if variants sometimes occurred at a lower frequency (but never higher) or at a higher frequency (but never lower) than another linked variant. Using the frequency of the "terminal" variants (frequency of variant at F, indicates frequency of clone C-E-F in absence of variant G), the frequency of the sub-lineages was tracked over time (Figure 38a). Interesting, only one of these clones was found in patient 28 (H-G-F-E-C), which again strongly suggests that a population bottleneck has occurred during transmission.

6.3.4. Parallel adaptation to the lung between patients

Parallel evolution, where traits or mutations occur independently at a frequency higher than could be explained by chance alone, has been observed between patients for other cystic fibrosis pathogens such as *B. dolosa* (Lieberman, Michel *et al*. 2011). Using a similar approach, the data was screened for the occurrence of at least two non-synonymous mutations in the same gene occurring independently in different patients. Out of a total of 107 SNPs, three genes in addition to one gene-pair (Table 7), were identified as having more than one independent non-synonymous SNP. By randomly introducing 107 non-synonymous SNPs *in silico* a 1000 times across the genome, and assuming that all positions across the genome evolve at the same rate, it was found that that this wouldn't be a pattern expected by chance alone (P= 0.001, Figure 39). When considering the 92 synonymous SNPs detected within patients, only two genes were found to have gained a SNP in two different patients, a pattern not significant under the same test. Two of the non-synonymous variants may be implicated in pathogenicity, including the PhoRP global regulator system which has been frequently recognized as important for both drug resistance and virulence in *M. tuberculosis* (Walters, Dubnau *et al*. 2006). Intriguingly, the accumulation of non-synonymous SNPs in the *phoR* gene was also recently observed in a *M. a. bolletii* infection of a cystic fibrosis patient (Kreutzfeldt, McAdam *et al*. 2013). *FolP2* is also of interest as it has been identified as a possible drug target of sulfonamides (Gengenbacher, Xu *et al*. 2008), which are sometimes used to treat *M. abscessus* infections. Mutations were also detected in the global transcriptional regulator *crp/fnr* which may have roles in oxygen or redox sensing and orchestrating bacterial stress

responses (Akhter, Yellaboina *et al*. 2008). It is particularly interesting that a gene orthologous to *crp*, associated with oxygen-related gene regulation, was also found to accumulate SNPs in *B. dolasa* (Lieberman, Michel *et al*. 2011)

Table 7 - Genes with evidence of convergent evolution between patients and clusters. Patients in parenthesis indicate the presence of a non-synonymous minority variant, with the rest being called as consensus SNPs. *One variant is found in both 31 and 22 due to result of transmission so is only counted once.

Gene	Patients	Number of SNPs (minority variants)	Encodes
MAB_0674/MAB_673	20, 15, (*31)*	3 (1)	PhoRP two component system
MAB_3675	20, 24	3	Probable succinate dehydrogenase, flavoprotein subunit SdhA
MAB_1345	7, 11, (*7, 30 x2)*	2 (3)	Probable dihydropteroate synthase 2 FolP2
MAB_0416c	4, 33 x2, 31, 22	4*	cAMP receptor protein

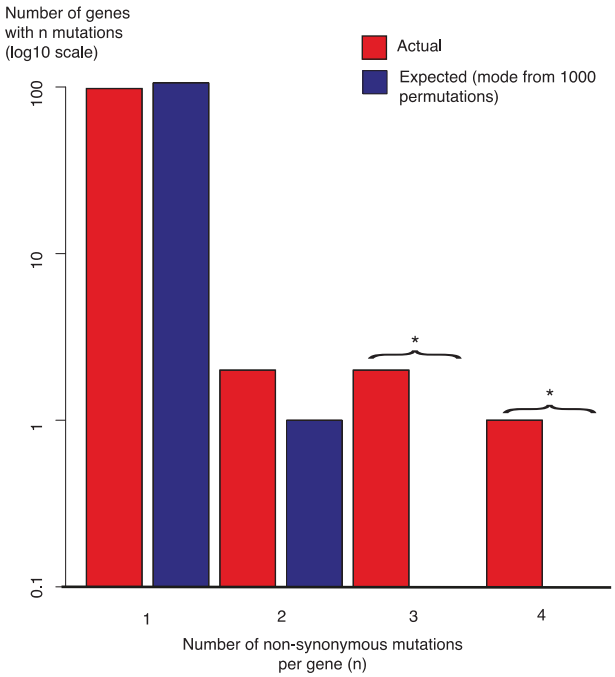


Figure 39 - Expected and actual number of independently acquired mutations per gene. To determine the expected frequency distribution of mutations within a gene, 107 non-synonymous mutations were introduced at random (the total number observed within patients in this study) *in silico* into the *M. abscessus* reference genome (Ripoll, Pasek *et al*. 2009)*;* this process was repeated a 1000 times. The average frequency of mutations per gene was counted (blue) and compared to the actual values (red) and plotted on a log scale. Four genes were found to have 3 or more non-synonymous mutations (Table 4), a pattern not expected by chance alone (stars indicate where the actual frequency

is observed in less than 5% of the 1000 permutations). It is important to note that this test is highly simplified in that it assumes that all positions across the genome mutate at the same rate.

6.3.5. Hypermutation

One patient's isolates were found to have a much higher number of minority variants, with an average of 35 per sample compared to rest of dataset with an average of 3.7 (as a result this patient was excluded from the longitudinal diversity analysis above). The excess number of these variants could not be explained by length of infection or treatment regimen. Both SNPs and minority variants were inspected manually, and a mutation conferring a premature stop codon in a uracil DNA glycosylase (*udg*) gene (MAB_3283c) was identified. *M. abscessus* has only one copy of the *udg* gene, which as part of the base excision repair pathway is responsible for removing miss-incorporated uracil from DNA. A knockout of *udg* in *Mycobacterium smegmatis* has previously been shown to confer a hypermutator phenotype with an excess of G/C->A/T transitions (Wanner, Castor *et al*. 2009). A similarly biased mutation spectrum was found for the minority variants detected in this patient (Figure 40a), supporting the inference that this mutation is having a similar effect in *M. abscessus*.

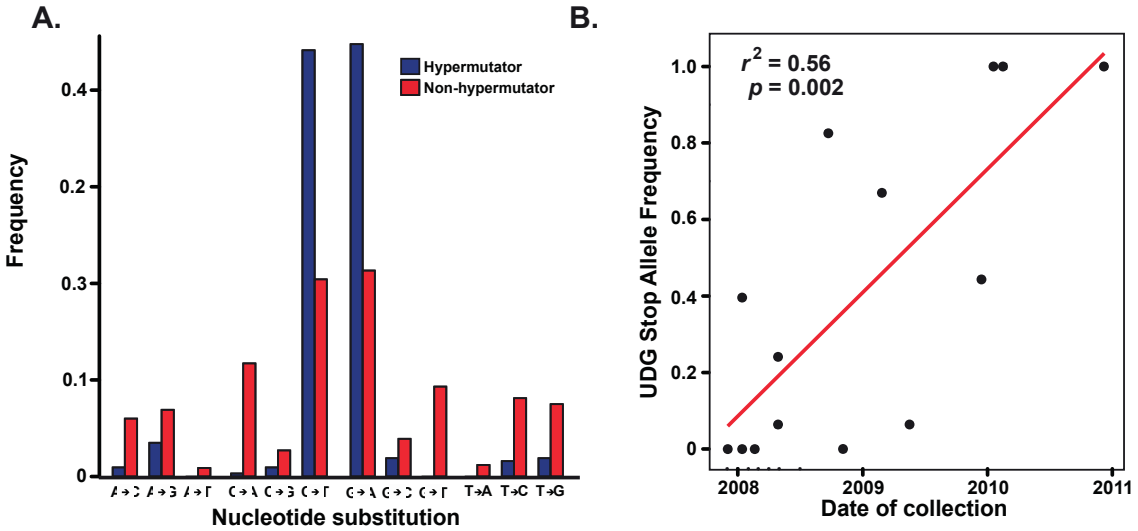


Figure 40 – Mutation frequency and spectra of the hypermutating strain. A) Mutation spectrum of minority variants from hupermutator (blue) and control (red) clones. Control frequency was obtained from patient 2 which has a similar number of isolates. The excess C to T and G to A nucleotide substitutions confirm functional loss of uracil DNA glycosylase (*udg*) in the hypermutator clone. B) Frequency of the hypermutator causing allele in the *udg* gene over time.

By using the depth of coverage that supported the *udg* mutation, the frequency of the hypermutator allele was found to increase over the course of the infection, reaching fixation in the most recent samples (Figure 40b). The evolutionary fitness of the hypermutator could not be ascribed to acquisition of another single gene mutation (as none exactly correlated in frequency with the *udg* mutator allele). However, it was striking that several genes were found to have accumulated non-synonymous SNPs, including the enhanced intracellular survival (*eis*) gene (MAB_4532c), which accumulated five mutations including a premature stop codon. This gene is considered important for virulence in *M. tuberculosis*, through suppression of host immune responses (Kim, An *et al*. 2012) and autophagy (Shin, Jeon *et al*. 2010). In addition, *nrdI* (MAB_3414c) accumulated three mutations, whose functional relevance is less clear; however this protein belongs to class 1 of ribonucleotide reductases, which in *E. coli* has been found to be activated by oxidative stress and iron limited conditions (Cotruvo and Stubbe 2008).

When considering the clinical impact of the hypermutator clone it was found that initially, infection with *M. abscessus* led to a rapid decline in lung function. However, fixation of the *udg* mutation within the patient in 2010 was associated with stabilization of both lung function and CRP with no further requirement for intravenous NTM therapy, suggesting a shift towards chronicity and an attenuation of virulence had occurred.

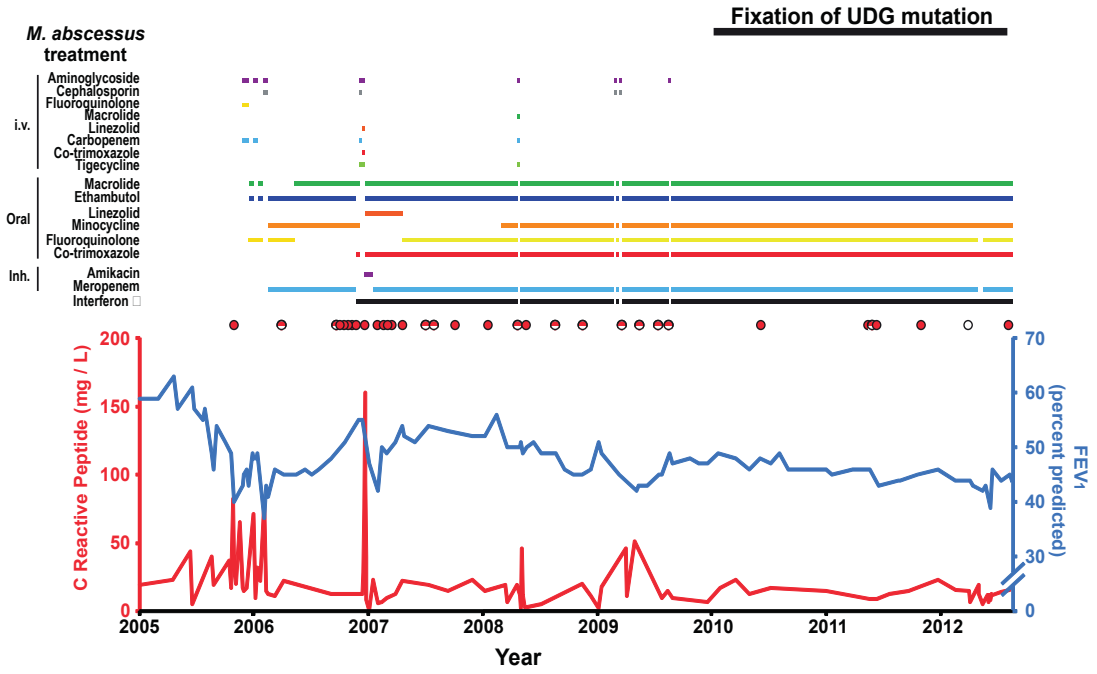


Figure 41 - Clinical trajectory of patient infected with hypermutator clone. Diagram shows changes over time in systemic inflammation (C Reactive protein; *red*), lung function (FEV1; *blue*), sputum cultures (smear negative and culture negative: *white circle*; smear negative and culture positive: *half red circle*; smear positive and culture positive: *red circle*) and intravenous (i.v.), oral and inhaled (*inh*) antibiotic therapy for *M. abscessus*. Fixation of *udg* mutation within the infecting *M. abscessus* population coincides with a more stable clinical course not requiring intravenous antibiotics

6.3.6. Functional consequences of genetic diversity

The presence of genetic heterogeneity in these samples would be expected to lead to functional heterogeneity. By evaluating 10 individual colonies from each *M. abscessus* sample using broth microdilution, considerable heterogeneity was observed in the minimum inhibitory concentration (MIC) of commonly used antibiotics (Figure 42). Corresponding to this phenotypic diversity, several polymorphisms were found in genes associated with antibiotic resistance (Table 8) including 16S rRNA (conferring aminoglycoside resistance (Prammananan, Sander *et al*. 1998)), penicillin binding proteins and beta lactamase (which may confer resistance to carbapenems and aztreonam (Yamachika, Sugihara *et al*.), (Liao and Hancock 1997), and the dihydropteroate synthase *folP2* (a potential target for sulphonamides (Gengenbacher,

Xu *et al*. 2008)). There was also substantial phenotypic heterogeneity in colony morphology (Figure 43), where switches occurred between smooth, rough and mixed morphotypes over the course of longitudinal sampling (examples in Figure 42). In some but not all cases, these switches could be associated with mutations in genes implicated in glycopeptidolipid synthesis, the pathway underlying this morphology (Table 9).

Table 8 - SNPs in antibiotic resistance loci generated within patients. Non synonymous variants in genes associated with antibiotic resistance were identified in patients isolates (the parenthesis indicates the specific sample).

Patient Sample†	Variant type	Gene function	Gene ID
1 (a)	Non-synonymous minority variant	Penicillin binding protein	MAB_3167c
22 (e)	Non-synonymous minority variant	Penicillin binding protein	MAB_3167c
1 (g)	Non-synonymous minority variant	Beta-lactamase	MAB_4947
5 *	A1408G	16s ribosomal RNA	MAB_r5051
9 (c)	A1408G	16s ribosomal RNA	MAB_r5051

†All samples, except 9 (c), were taken from patients established on treatment for *M. abscessus*.
* most patient 5 isolates have a level of heterogeneity at this position

Table 9 - SNPs in GPL loci. Non synonymous or frameshift variants were identified in the GPL loci genes (Ripoll, Deshayes *et al*. 2007) in patients where a colony morphology switch or heterogeneity was observed.

Patient	Variant type	Gene	Phenotype
19 (n)	Non-synonymous minority variant	MAB_0939	S/R
5 (m)	Frameshift insertion	MAB_4098c	R
5 (j)	Heterogeneous (30%) frameshift deletion	MAB_4099c	S/R
1	Frameshift insertion	MAB_4099	S/R

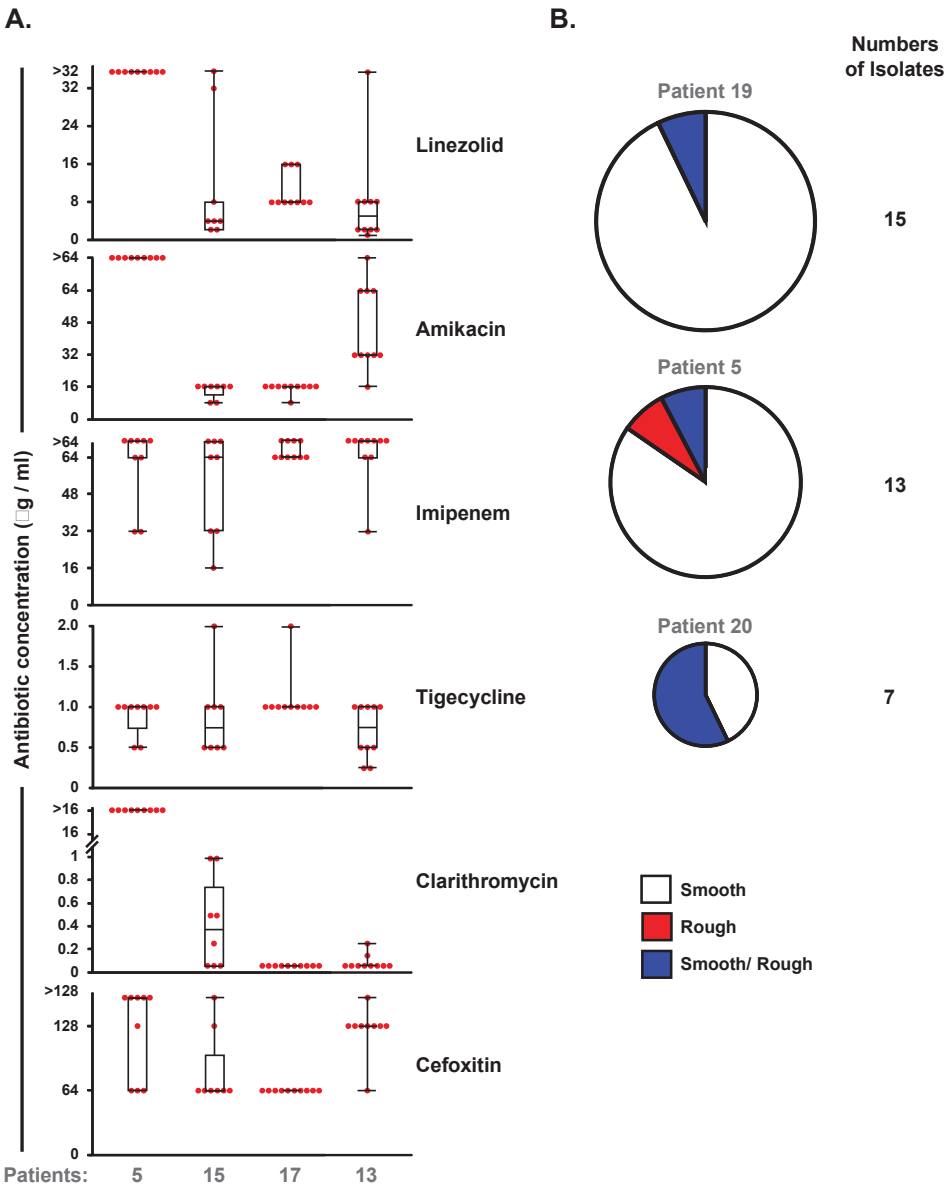


Figure 42 - Within patient phenotypic heterogeneity. A) Within sample variation in antibiotic susceptibility measured by broth microdilution in four patients and 10 colonies. B) Variation in colony morphotype of isolates from four individuals chronically infected with *M. abscessus* was established by growing samples on solid media and inspecting whether colonies were smooth (white), rough (red) or a combination of smooth and rough (blue).

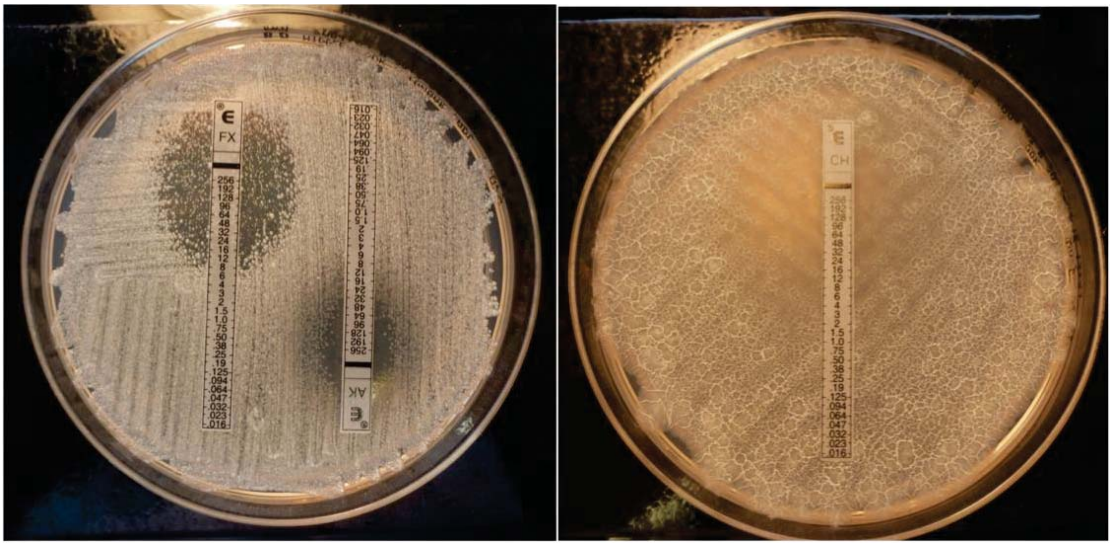


Figure 43 – Smooth (L) and rough (R) colony morphotypes of *M. abscessus*. Picture provided by Dorothy Grogono.

6.4. Discussion

Traditional genotyping and whole genome sequencing techniques of bacteria have often relied on the isolation of a single colony, and this single isolate is assumed to represent the infecting population. It is increasingly being appreciated however, that significant within patient diversity can exist across a wide range of host environments and time scales as shown for *Staphylococcus aureus* (Golubchik, Batty et al. 2013, Harris, Cartwright et al. 2013), *Helicobacter pylori* (Kennemann, Didelot *et al*. 2011) and *M. tuberculosis* (Perez-Lago, Comas *et al*. 2014). This has been more intensively explored in cystic fibrosis pathogens such as *P. aeruginosa* (Smith, Buckley et al. 2006, Chung, Becq et al. 2012) and *B. dolasa* (Lieberman, Michel et al. 2011, Lieberman, Flett et al. 2014), due to the chronicity of their infections. However, how this diversity relates to clinical phenotype, and its impact on how we understand within-patient evolution and transmission is yet to be fully explored.

As predicted by population genetics, this study provides evidence that genetic diversity correlates with population size and that this population size fluctuates over the course of infection. In asymptomatic *S. aureus* carriage, similar fluctuations have also been observed (Golubchik, Batty *et al*. 2013), which are assumed to relate to

cycles in colonization and clearance. In *M. abscessus*, these fluctuations frequently correspond to changes in the clinical phenotype with explosions in genetic diversity coinciding with infective exacerbations and conversely population bottlenecks occurring after intensification of antibiotic treatment. This is the first study, to my knowledge, which has been able to relate changes in within-host diversity of a bacterial pathogen to clinical phenotype and treatment.

A population bottleneck appeared to occur upon transmission of *M. abscessus* from one patient to another, suggesting a limited infectious dose. Although this study is limited to one transmission event, *M. abscessus* appears to be similar to *M. tuberculosis*, where even though significant within-patient diversity is generated, the heterogeneity appears to be lost upon transmission, with only one genotype being found in the recipient patient (Perez-Lago, Comas *et al*. 2014).

This study found that clones picked from a single sample can have different antibiotic profiles. The current microbiology standard is to perform broth microdilution, which can be performed on a clonal culture or a sweep. However both of these are likely to be unrepresentative of the overall population. The former provides results for a single organism, which may not be typical of the rest of the population, whilst the latter is biased towards resistant organisms which will grow even if they are present only as a minority. Antibiotic susceptibility testing results performed in this way are therefore likely to be misleading, and may explain why current results are often inconsistent with patient outcome and can be irreproducible (Broda, Jebbari *et al*. 2013).

Whilst much of this genetic diversity reflects neutral variation, some of it appears to be associated with the observed phenotypic diversity, so could have a selective advantage. In addition, parallel evolution between patients was observed, where several genes were found to accumulate non-synonymous variants independently in different patients; many of which have been associated with virulence or drug resistance in *M. tuberculosis*. In addition to this, several variants in one chronically infected individual were found to change frequency over time and showed linkage to one another. Reconstruction of haplotypes from minority variants has only previously been achievable with longer reads for short viral genomes such as HIV (for example (Zagordi, Klein *et al*. 2010)) and cancer (Fischer, Vazquez-Garcia *et al*. 2014).

However, due to the dense longitudinal sampling, this was able to be achieved with short un-linked reads by correlating them over time, a pattern previously noted for the influenza virus (Watson, Welkers *et al*. 2013). For *M. abscessus* this revealed a pattern of linked lineages that are associated with non-synonymous variants in genes associated with host adaptation or virulence. This suggests that this population structure has been selected for through a series of bottlenecks or selective sweeps. However, the lineages appear to be present across the sampling period, suggesting a lack of replacement by one lineage over another. This has previously been observed in *B. dolosa*, where several adaptive lineages have been observed to coexist over time, with a lack of fixation (Lieberman, Flett *et al*. 2014). This could be explained by geographical separation of different lineages within the lung niche, allowing local selective sweeps. This compartmentalisation has strong parallels with *M. tuberculosis*, which forms multiple discrete lesions in the lung tissue. Although it is currently not known whether *M. abscessus* forms granulomatous lesions in cystic fibrosis patients, the pattern we see in this chronically infected patient is highly suggestive of this kind of geographical structuring.

Finally, this study identified the first clinical hypermutator for a mycobacterium. The hypermutator accumulated many mutations in virulence-associated genes; but some of these resulted in a loss of function, indicating that these mutations are resulting in attenuation of *M. abscessus*. The fixation of the hypermutator was not associated with a clinical decline, indicating that the presence of the hypermutator hadn't worsened the patient's symptoms. These analyses together support a scenario where selection is occurring within patients for variants that favor long term survival and adaptation of *M. abscessus* to the cystic fibrosis lung niche, and a shift away from acute to more chronic infections. Hypermutators provide us with an opportunity to observe an accelerated version of the evolutionary trajectory of bacterial adaptation to the host, and it was found that the same genes accumulated mutations in non-hypermutator patients also. This suggests that the hypermutator's evolutionary trajectory towards attenuation may be shared by other *M. abscessus* infections, and also by other cystic fibrosis species, as this trend has been observed for *P. auriginosa* (Smith, Buckley *et al*. 2006). This, in addition to other parallels such as colony morphology switches (Govan and Deretic 1996) and convergent evolution, suggest that very different cystic

fibrosis pathogens have highly similar ways of adapting to the cystic fibrosis lung niche.

7. Conclusions

7.1. A restatement of the research questions and aims

Mycobacteria are characterised by slow growth and therefore a slow mutation rate, which makes the application of whole genome sequencing, and the resolution it can provide, particularly attractive. The aim of this thesis was to use whole genome sequencing to understand the evolutionary dynamics and transmission of two mycobacterial species: *M. tuberculosis* and *M. abscessus*. *M tuberculosis*, as a major cause of infectious disease globally, is one of the most well studied bacterial pathogens in microbiology. Despite this there are still aspects of its biology and transmission that are poorly understood. Primarily this thesis is concerned with its genome-wide genetic turnover in the context of transmission and recurrent disease. For *M. abscessus*, an emerging cystic fibrosis pathogen, very little was known concerning its population structure and genome-wide diversity before the initiation of this study, with only one available genome sequence. Assumptions had been made on its mode of transmission, based on a handful of studies using genotyping techniques that lack resolution such as MLST (Macheras, Roux *et al*. 2011) or PFGE (Bange, Brown *et al*. 2001). This study aimed to capture its population structure, in order to improve our understanding of the way it is acquired by patients, and the diversity within patients.

7.2. Key findings

7.2.1. The molecular clock of mycobacteria is slow, which impacts on how we interpret its application to transmission or recurrent disease

M. tuberculosis was estimated to have an average substitution rate of 0.3 SNPs per genome per year. In practice this means two isolates collected three years apart could be expected to be identical, and that a close genetic distance cannot be used alone to infer direct transmission. Supporting this, another study found that there was a poor correlation between geographic and genetic distance in Samara, Russia (Casali, Nikolayevskyy *et al*. 2014), with identical isolates being isolated from patients living over 125km apart. *M. abscessus* was found to have a slightly higher genetic turnover, at 0.5-1.8 SNPs per genome per year depending on the subspecies. Even this small increase in the molecular clock made analyses and interpretation possible (such as

coalescent analyses) that weren't possible with the *M. tuberculosis* datasets. However these rates are still low enough to lead to misinterpretation and inferences of transmission events that are extremely unlikely, such as a direct cross-continental transmission as suggested for two *M. abscessus* outbreaks in Papworth and Seattle (Tettelin H, Davidson R.M *et al*. 2014), in the absence of any other supporting evidence.

Due to this low rate of genetic turnover, other pieces of evidence become more important when investigating transmission. Epidemiological evidence, in the form of contact tracing, is still vital. For *M. abscessus* the overlap in patient visits was used for not only demonstrating that transmission was possible for the patients with clustered isolates, but also revealed that the mode of transmission was unlikely to be direct.

Another tool when interpreting transmission is the topology of the phylogeny. In this thesis, the phylogeny was used to make an assessment on the likelihood of transmission for both *M. tuberculosis* and *M. abscessus*. For *M. tuberculosis*, with only one isolate per patient, the phylogeny itself was unable to confirm transmission, but instead could be used to exclude the possibility of direct transmission between two patients due to the positioning of other patients' isolates on intermediate nodes. For *M. abscessus*, multiple isolates per patient were available, meaning that overlaps in their diversity or inheritance of subsets of diversity could be detected, which provided strong evidence of transmission between some patients. These both highlight the importance of phylogenetic context, provided by multiple isolates from the same patient, in addition to isolates from other patients. The conclusion of patient-patient transmission for *M. abscessus* would have been strengthened by the availability of environmental isolates, which would have provided the phylogenetic context to the population structure derived from clinical isolates and also would have allowed a more confident assessment of the possibility of an environmental point-source. In addition, intensive sampling from other cystic fibrosis clinics would also provide context from presumably unrelated patients. However, it's often not possible to sample as intensively as desired, so for both the studies presented here and for future ones, a full consideration of the limits of sampling is essential when interpreting transmission.

For determining whether relapse or re-infection has caused a secondary disease episode, all the above considerations still apply with the exception of epidemiology in the form of contact tracing. If the patient has relapsed, the isolates would be expected to have a small genetic distance (in this study <=6 SNPs) and be sister taxa who share a most recent common ancestor on the phylogenetic tree. Re-infection cases however, should be phylogenetically disparate, lacking a most recent common ancestor or with a very large genetic distance (in this study >1000 SNPs) There is the possibility that apparent relapse cases could in fact represent re-infection with a highly related strain (from a family member for example), but there is no way to distinguish the two scenarios with a molecular clock is as slow as this. Again this demonstrates that although whole genome sequencing is useful, its limits for slow evolving pathogens need to be fully appreciated.

7.2.2. The PE and PPE genes of M. tuberculosis are not hyper-variable within patients

When comparing complete genome sequences of *M. tuberculosis* the PE and PPE genes have been documented as being hyper-variable. This hyper-variability is assumed to be due to *de novo* SNPs, insertions, deletions and recombination. Several studies (McEvoy, Cloete et al. 2012, Copin, Coscolla et al. 2014), have attempted to quantify the variation in the PE and PPE genes, but none have placed this variation in the context of phylogenetic scale. By placing the variation in the phylogenetic context of the Beijing lineage of *M. tuberculosis*, it was found that the pattern of variation introduced via *de novo* SNPs was indistinguishable from the rest of the genome, but that there was a higher rate of in-frame insertions and deletions. It is still unclear what the significance of this is, but further studies on different phylogenetic scales or even different Mycobacterial species (*M. abscessus* has only 9 PE or PPE genes annotated) may provide greater clarity.

7.2.3. Within host diversity of Mycobacterial infections can be the result of mixed infections or on-going evolution

Typically, whole genome sequencing of bacteria is carried out on a culture derived from a single colony. This has its advantages in that it simplifies phenotyping and interpretation of sequence data where only consensus variants need to be considered. However, a major disadvantage of this approach is that much of the diversity within a patient or the system being studied is left un-sampled. This may have resulted in an under-appreciation of mixed infections. When techniques are orientated towards the entire sample, rather than a single colony, a surprisingly high frequency of mixed infections of tuberculosis has been found (Warren, Victor *et al*. 2004). However these can be hard to interpret or distinguish from artifactual error when using standard genotyping techniques. Using whole genome sequencing however, it was found that a mixed infection could not only be detected (even when one strain was as low as 8%), but they could be disentangled using the phylogenetic context of other unrelated strains.

In addition to the within-patient diversity caused by multiple infections, there is also within-patient diversity generated by a continually evolving clonal infection. This has been demonstrated for *M. tuberculosis* (Sun, Luo *et al*. 2012), where between 8 and 41 variants have been found to be present within each of three patients. There was little evidence for this in this study, which is probably due to the microbiological methods being oriented towards a single colony, and only one or two samples being available per patient. For *M. abscessus* however, there was a greater opportunity to capture this diversity, as no colony purification had been carried out, and up to 28 samples per patient had been collected over the course of 4 years. This enabled significant diversity to be captured and quantified over the course of an infection, and related back to clinical phenotype and patient outcome in a way that hadn't been achieved for bacteria before.

However a major limitation of both of these investigations is that they relied on culture. It is unknown how much culture biases the resultant mixture that is sequenced but studies using PCR (Hanekom, Streicher *et al*. 2013), and also the finding that the mixed XDR infection was only detected using one culture technique but not the other, suggests that the potential impact could be large. In addition, the sample itself may be biased in that mycobacterial infections form local lesions, and sputum samples may only represent one or some of these lesions at any one time. New experimental

approaches will be required in the future to overcome this, or better quantify the impact of sampling in different infectious disease contexts.

7.3. Clinical impact

7.3.1. *M. abscessus* is able to transmit between cystic fibrosis patients

The finding that *M. abscessus* was able to transmit between patients in a cystic fibrosis clinic has multiple clinical implications. Firstly it means that infection control procedures need to include *M. abscessus* in addition to the other pathogens known to transmit. As a direct result of this study and others (Aitken, Limaye *et al*. 2012), guidelines concerning *M. abscessus* are currently being amended in the UK and the US (verbal communication – Andres Floto). However, in order to make more specific recommendations, knowledge of the route of transmission will need to be better characterised. On the analysis presented here the most likely mechanism appears to be an indirect one, that could be due to either aerosols or the shedding of fomites. Determining this will be really important for implementing the most effective infectious control measures.

A second implication of this work, is that it highlights how little we understand about patient-patient transmission of cystic fibrosis pathogens, as although a high rate of transmission was found for *M. abscessus* in Papworth, the hospital has very low rates of *P. aeruginosa* and *Burkholderia cepacia* complex transmission. This suggests that the different pathogens have different mechanisms of spread. Differing rates of transmission might also be due to geographic spread of transmissible clones, as the rate of transmission might be dependent on the opportunity to be infected by a transmissible clone in the first place. *M. abscessus* incidence is known to vary with geography (Hoefsloot, van Ingen et al. 2013, Chou, Clements et al. 2014), suggesting that both environmental and human reservoirs need to be considered in order to understand this system fully.

Finally this work also highlights the possibility that other under-studied cystic fibrosis pathogens may also be able to transmit, and that it may be dangerous to assume otherwise.

7.3.2. Whole genome sequencing of Mycobacteria in the clinic: tracking transmission

It is inevitable that in the near-future clinical *M. tuberculosis* isolates will be whole genome sequenced routinely for the purpose of outbreak investigation, and this is something Public Health England has included in their new strategy (Public Health England 2014). In order for this to be feasible, approaches need to be developed to allow infection control teams to interpret the genetic distances that whole genome sequencing will provide. Its been proposed that a simple threshold such as five SNPs for samples isolated less than three years apart might be appropriate (Walker, Ip *et al*. 2013). However the work presented here suggests that such a simple cutoff might not be appropriate, as although 0.3 SNPs per genome per year is the average rate of change, there is a lot of variability around it. Instead, phylogenetic context proved to be more useful as discussed above. In addition it is currently not known whether *M. tuberculosis* hypermutators exist, although this study did demonstrate that this is the case for *M. abscessus*, which could further obscure any thresholds used. Instead it is likely that phylogenetic trees and genetic distances derived from whole genome data will need to be used as tools and pieces of evidence as part of an over-arching judgment based on information from several sources. Only experience, informed by larger scale studies (for example in Oxfordshire (Walker, Lalor *et al*. 2014)) will enable these systems to develop.

7.3.3. Whole genome sequencing of Mycobacteria in the clinic: antibiotic resistance

All currently described antibiotic resistance mechanisms in *M. tuberculosis* are chromosomally encoded, meaning that sequence based tests are easy and simple to interpret. This has led to the success of nucleic acid amplification based tests such as the GeneXpert (Helb, Jones *et al*. 2010). Whole genome sequencing is an attractive

alternative to this system as multiple loci can be detected in one "test", and also has the ability to detect plasmids which may encode resistance genes, as demonstrated for *M. abscessus* (Matsumoto, Bispo *et al*. 2014). However, this thesis has raised a number of issues that will need to be considered if this is implemented.

Firstly, mixed infections and the diversity of clonal infections could result in minority or mixed resistance phenotypes. This was found for the patient with an XDR infection, which was comprised of two XDR strains with independently acquired resistance. In this case the mixture was 70:30, but in other cases there may be a minority strain at a very low prevalence. For *M. abscessus* there were situations where resistance was acquired during the course of infection, so could be at a minority. Systems would need to be put in place to detect these scenarios.

Secondly, it is unknown how much resistance is currently un-described. This is something that could not be estimated in this thesis due to the vast majority of *M. tuberculosis* isolates being fully sensitive, however in the context of Samara, Russia it was found that the vast majority of resistance could be explained by known mechanisms (Casali, Nikolayevskyy *et al*. 2012), however it is unknown how much could be explained in other contexts. For *M. abscessus* several variants were identified that occurred in possible drug-targets or associated genes, that haven't been described before. These would need to be validated, but highlight how poorly this is currently described. Clinical based systems utilising mcobacterial genomes would need to take account of these limitations and ideally new resistance loci could be "learnt" iteratively as the databases grow.

7.4. Future directions

The field of microbial genomics is rapidly expanding, meaning that sample collections are constantly getting larger and the technology is getting better. Future work on mycobacteria will build upon some of the findings presented here and will hopefully refine our knowledge regarding transmission and evolutionary dynamics. In particular there are a number of aspects of this thesis that open up further questions for investigation.

Firstly, the observed population structure of *M. abscessus* suggests the presence of dominant circulating clones. In order to confirm the existence of these clones further sequencing of collections from outside Papworth, and outside cystic fibrosis patients is required. This would enable us to understand their reach and nature, in addition to providing insights into the genetic basis of their success. To this end a global collection of over 1,700 *M. abscessus* isolates are currently being sequenced at the WT Sanger Institute, which will hopefully enable us to answer these questions. However this collection only contains a small number of environmental isolates, so more will need to be collected in order to understand the population structure of the environmental reservoir of *M. abscessus* to provide context to the clinically derived isolates.

This thesis involved an investigation of *M. tuberculosis* diversity over three scales: at the lineage, transmission and patient level. However, there may be more scope for investigating the patient level, as this thesis only involved two isolates per patient at most. The investigation into the within-patient diversity of *M. abscessus* revealed a high level of diversity, which fluctuated over time. This kind of in-depth analysis hasn't been carried out for *M. tuberculosis*, so it is currently unknown how much diversity exists, although one small-scale study suggests it can potentially be quite high (Sun, Luo *et al*. 2012). This could not only increase our understanding of how *M. tuberculosis* diversity relates to time and space within a patient, but also would have clinical relevance in terms of antibiotic resistance, and allow us to observe how it evolves in real-time. For this kind of study to be truly representative of a patient's infection, the limits of culture and clinical sampling would need to be overcome. Deep sequencing without the requirement of culture would enable the sample's diversity to be properly represented, and one way to achieve that could be through nucleic acid capture techniques (Depledge, Palser *et al*. 2011). Currently, nearly all clinical samples of mycobacteria are derived from sputum which is likely to be extremely biased in terms of what lesion is discharging into the airways. So in order to overcome the limitations of sputum, sampling would need to be carried out via autopsy or from transplanted lungs. This would also allow the sampling of multiple pathogens at once, such as *M. tuberculosis* and HIV, which are well known to co-infect. A multi-pathogen approach would be particularly important in the context of cystic fibrosis where many of the clinical phenotypes that were correlated with *M.*

abscessus diversity, could have been obscured by changes in the burden of additional co-infecting pathogens. Sampling of the multiple pathogens in cystic fibrosis patients at once would give us greater insight into the entire system and its impact on clinical outcome.

Finally, in addition to the two species studied in this thesis, there are many other members of the genus Mycobacteria that are pathogenic to humans and animals. Many of the same principles presented here could be applied to them. In particular the transmission route of *M. ulcerans* is still not known, although it is suspected to transmit to humans through aquatic biting insects (Johnson, Stinear *et al*. 2005). With an even slower growth rate than *M. tuberculosis,* whole genome sequencing rather than traditional genotyping will be required to understand this system better. In the UK, there is high concern regarding *M. bovis,* which is responsible for a very large burden of disease in cattle. Sequencing of both the cattle and wildlife reservoirs (in particular badgers) holds great promise for learning how to tackle this disease.

7.5. Closing comments

Whole genome sequencing has revealed the evolution of two important mycobacterial pathogens over different evolutionary scales including the patient, transmission and species levels. These analyses have not only informed us how they evolve, and at what rate, but also have had a significant clinical impact. More generally, they provide a framework for how whole genome sequencing can be used to provide us with insights into transmission and evolutionary dynamics of pathogens, particularly those with slow molecular clocks.

8. Methods

This section includes all bioinformatics methods used in this thesis. Laboratory techniques and experiments carried out by others are not included. Many of the programs described below form part of in-house scripts and pipelines made available for use by the informatics team and group members at the WT Sanger Institute, and are stated as such.

8.1. Illumina sequencing

The DNA pipeline teams at the WT Sanger Institute carried out all library preparation and sequencing unless stated otherwise. All sequencing was carried out in a paired end and multiplex (12-96 samples per run) fashion on the GAIIx, HiSeq2000 or MiSeq platforms.

8.2. Mapping of sequencing data to a reference sequence

Paired-end reads (fastq file format (Cock, Fields *et al*. 2010)) were mapped to a suitable reference (Table 10) using the program SMALT (v0.5.8) (Ponstingl 2011) uniquely (reads with multiple best matches are discarded) using default parameters except for:

- Maximum insert size of 1000 (-i)
- Minimum insert size of 50 (-j)
- Turn on exhaustive search were each mate of paired end reads are mapped independently (-x)
- Filter out aligned reads that don't have a certain proportion of exact matches (-y – used default of 0 unless stated otherwise)

Table 10 - Description of reference sequences used for mapping analyses. *Illumina reads were generated at a later date and were used to iteratively correct the reference using iCORN which resulted in 32 single base pair corrections.

Species	**Strain name**	**Type**	**Accession or source**
M. tuberculosis	H37Rv	Finished genome (AL123456) with corrections*.	Casali, Nikolayevskyy *et al*. 2012
M. tuberculosis	2535G	PacBio assembly	Currently unavailable
M. abscessus	CU458896	Finished	NC_010394.1
M. abscessus	3_k, subspecies *bolletii* representative	*De novo* assembly of Illumina reads	ERR115028 (raw reads)
M. abscessus	22_e, subspecies *massiliense* representative	*De novo* assembly of Illumina reads	ERR115082 (raw reads)

8.3. Calling and filtering variants from mapping data

Samtools and bcftools (Li, Handsaker *et al*. 2009) were used to call bases as part of an in-house pipeline (written by Simon Harris). Appropriate filters were used to reduce the number of false positive SNP calls to a level estimated to be less than one SNP per genome (Harris, Feil *et al*. 2010). Filters were:

- The minimum base quality to call a base is 50 (Phred score)
- The minimum mapping quality to call a base is 30
- The minimum number of high quality reads mapping to call a base is 4.
- The minimum number of high quality reads mapping on each strand to call a base is 2.
- The minimum proportion of high quality mapped reads that must match the called base is 0.75
- Minimum P value for strand bias, base quality bias, mapping quality, and end bias of 0.001

Any positions that failed the quality criteria were called as 'N' in the final alignment.

8.4. *De novo* assembly of sequencing reads

Raw sequencing reads were assembled using Velvet v1.2.03 (Zerbino and Birney 2008) using an in-house script which optimises the k-mer (hash length) with an expected depth of coverage of 20.

8.5. Multiple sequence alignment

Sequences (FASTA format) were aligned using Muscle v3.8.31 (Edgar 2004) or MAFFT vs. 7 (Katoh and Standley 2013) with default settings.

8.6. Construction of maximum likelihood phylogenetic trees

Maximum likelihood phylogenetic trees were constructed using RAxML v. 7.0.4 (Stamatakis 2006) using the GTR+GAMMA method for among site rate variation and 100 bootstrap replicates. The final tree was created using the maximum likelihood

tree topology with the calculated bootstrap values drawn onto the bipartitions. In some cases it was useful to map the variants back onto the resultant tree topology in order to scale the branch lengths by number of SNPs. This was carried out using an in-house script (written by Simon Harris) which utilised ACCTRAN parsimony algorithms (Farris 1970).

8.7. Path-O-Gen analysis

Path-o-gen (Rambaut 2007) is a program used to assess the presence of molecular clock in a dataset. Using a phylogenetic (often maximum likelihood) tree as input, it plots the relationship between root-to-tip distance for each taxa with its isolation date. Under the assumption of a strict molecular clock there should be a strong positive linear relationship between root-to-tip distance and time. The program can be used to root the tree in the position most consistent with this relationship. This power of this analysis is limited by its non-independence as many taxa will share evolutionary history and therefore their root-to-tip distance will not be independent. Therefore where this has been used, P values have not been stated.

8.8. Bayesian molecular evolution analysis

The BEAST package (v1.7.5), a program used for Bayesian Markov chain Monte Carlo (MCMC) analysis of genetic sequences, was used to estimate the mutation rate and the age of phylogenetic nodes (Drummond and Rambaut 2007). BEAST requires XML files as input where all the priors are set, which were created was using the GUI BEAUTi (Drummond and Rambaut 2007) and nucleotide alignments and the associated dates of isolation as input. For all analyses three independent MCMC chains of 100,000,000 states were run using a GTR model of evolution and a variety of different clock and population size models. Tracer (v1.5) (Drummond and Rambaut 2007) was used to assess convergence (after an initial burn-in period of 10,000,000), agreement between the three runs and that all effective sample size (ESS) values were greater than 200. When the uncorrelated lognormal relaxed clock was used, there was no appreciable probability mass in the marginal posterior distribution of the standard deviation of the clock rate (ucld.stdev) that overlapped with zero, so a strict clock was not deemed appropriate. For each dataset tested, one

run with the best ESS values was used to produce a maximum clade credibility tree in TreeAnnotator v1.7.1 (Drummond and Rambaut 2007), from which the estimated age (and the 95% higher posterior density intervals) of the internal nodes were extracted.

8.9. Statistical analyses and figures

Statistical analyses were implemented in R version 3.0.0 (R Core Team 2013). Figures generated using R or Microsoft Excel version 14.3.9 (Microsoft Corporation 2011).

8.10. Detection of heterogeneous sites / minority variants

Many heterogeneous or minority variants found in mapping data are due to sequencing error or mismapping, therefore strict filters are required to distinguish these errors from true minority variants which are the result of a mixed infection or clonal variation. After mapping to a reference, positions where two or more possible variants were called were extracted, and then each variant had to pass certain quality criteria:

- Supported by at least two reads on both forward and reverse strands
- Overall quality greater than 100
- Minimum P value for strand bias of 0.05
- Minimum P value for base quality bias, mapping quality, and end bias of 0.001
- Depth of coverage within normal range (+-50% of the average calculated from bam file)
- At least 200bp from another variant.

These criteria were chosen on the basis that these are the requirements for calling consensus SNPs (section 8.3), however it was found that the strand bias and mapping parameters were required to be more stringent when considering minority variants. Using the standard mapping parameters (see section 8.2), many of the detected

minority variants were found in GC rich regions clustered together (Figure 44). These were due to sequence specific errors that frequently appear after a GGC/GCC motif in Illumina reads (Nakamura, Oshima *et al*. 2011). In addition the mutation spectrum of the minority variants detected closely matched the error profile of these regions (Figure 45).

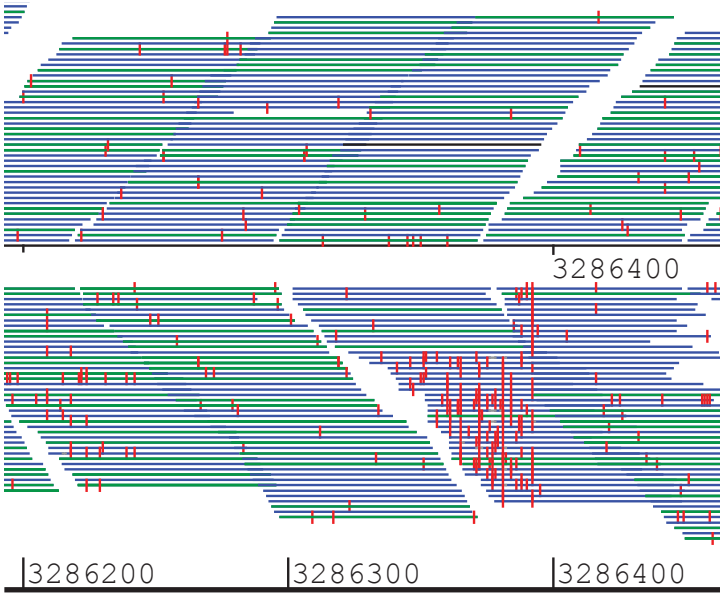
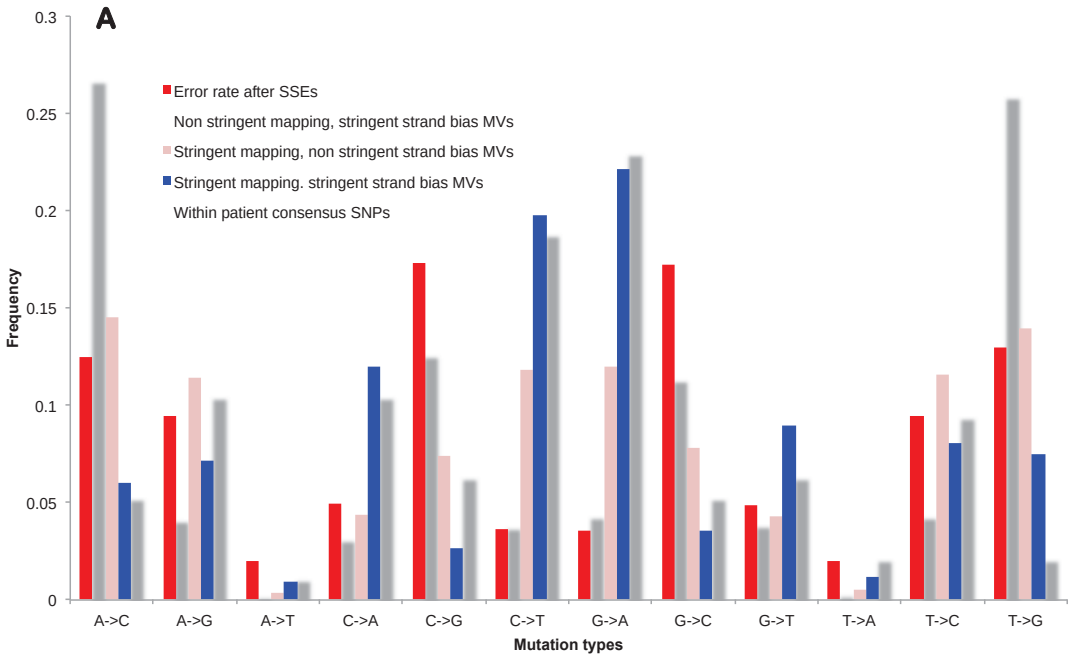


Figure 44 - Example of sequencing errors that occur after GGC/GCC motifs on the reverse strand. Each read is represented by a blue or green line, which are split into those that map to the forward and reverse strands. Red marks indicate bases which have been called as a different nucleotide to the reference genome. Numbers along the bottom indicate the position on the *M. abscessus* (CU458896) reference. Strand specific errors are found clustered between 3266343 and 3286392. Figure made and adapted from Artemis.

In order to avoid these variants, the stringency of the mapping was improved by using a 0.95 value for minimum match (section 8.2). This means any reads of length 75bp with 4 or more mismatches were discarded. In addition it was found that an increased stringency for strand bias was required, with a minimum P value of 0.05 rather than 0.001. In combination, these two additional parameters resulted in a mutation spectrum that highly resembled the mutation spectra of consensus SNPs (Figure 45 - blue).



B

	Error rate SSEs	Non stringent mapping	Stringent mapping, non stringent strand bias	Stringent mapping, stringent strand bias
Error rate SSEs				
Non stringent mapping	0.69			
Stringent mapping, non stringent strand bias	0.45	0.63		
Stringent mapping, stringent strand bias	-0.36	-0.15	0.47	
Consensus SNPs	-0.28	-0.29	0.42	0.93

Figure 45 - Mutation spectra of minority and consensus variants. A) Sequence specific error (SSE) profile obtained from Nakamura *et al* 2011. Other mutation spectra were obtained from *M. a. massiliense* dataset comprised of the two transmission clusters. Stringent and non stringent parameters are described in the text. MV = minority variant. **B)** Pearson's R coefficient values between mutation spectra, with colours representing the strength of the relationship. The mutation spectra of the consensus SNPs was found to correlate highly with the minority variants detected through stringent mapping and stringent strand bias.

Variants were also only included if they were at least 200bp from another possible minority variant; this was to exclude heterogeneity which may be the result of mismapping. This distance may need to be reduced in non-mycobacterial organisms where the mutation rate is sufficiently high enough for mutations to frequently occur within 200bp of one another.

The number of minority variants detected was found to significantly correlate with the depth of coverage (Figure 46), which is expected as deeper sequencing would increase the ability to both detect minority variants and for them to subsequently pass quality filters. Therefore, when comparing the number of minority variants between different samples, the data was normalised to 100 fold depth of coverage using the following calculation:

$$corrected\ no.of\ minority\ variants = \frac{no.of\ minority\ variants}{average\ depth\ of\ coverage} \times 100$$

This normalisation resulted in a loss of any observable correlation between the two variables (P value 0.2, correlation coefficient 0.02).

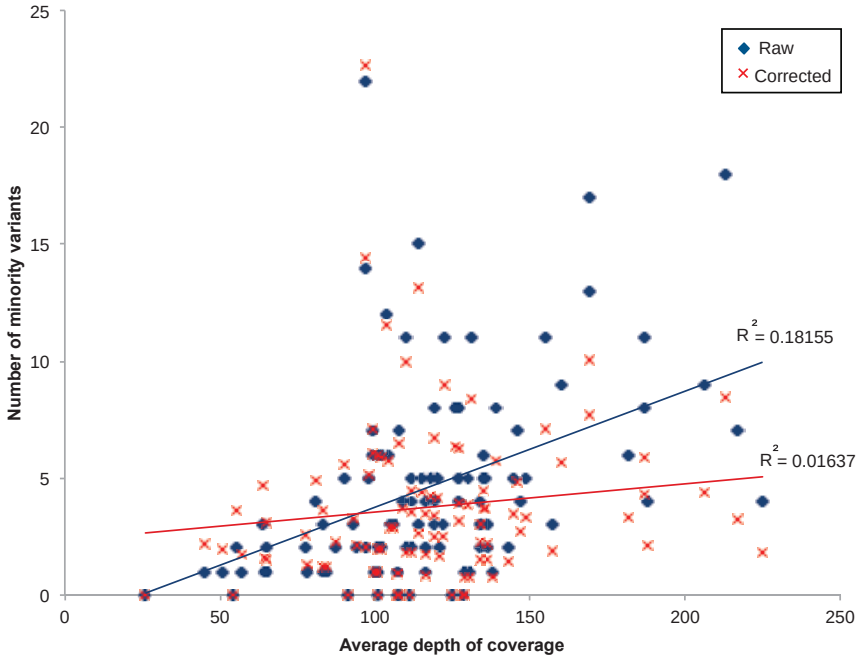


Figure 46 - Correcting the number of minority variants for depth of coverage. Analysis of the *M. abscessus* dataset (Chapter 6, excluding the hypermutator) revealed a highly significant (P value 8.893e-06) positive linear relationship between the raw number of minority variants detected and coverage (blue). When normalised for coverage (red) no significant correlation could be observed (P of 0.2).

10. References

Abdallah, A. M., T. Verboom, E. M. Weerdenburg, N. C. Gey van Pittius, P. W. Mahasha, C. Jimenez, M. Parra, N. Cadieux, M. J. Brennan, B. J. Appelmelk and W. Bitter (2009). "PPE and PE_PGRS proteins of Mycobacterium marinum are transported via the type VII secretion system ESX-5." Molecular Microbiology **73**(3): 329-340.

Abebe, F. and G. Bjune (2006). "The emergence of Beijing family genotypes of Mycobacterium tuberculosis and low-level protection by bacille Calmette-Guerin (BCG) vaccines: is there a link?" Clinical and Experimental Immunology **145**(3): 389-397.

Adekambi, T., M. Reynaud-Gaubert, G. Greub, M. J. Gevaudan, B. La Scola, D. Raoult and M. Drancourt (2004). "Amoebal coculture of "Mycobacterium massiliense" sp. nov. from the sputum of a patient with hemoptoic pneumonia." Journal of Clinical Microbiology **42**(12): 5493-5501.

Aitken, M. L., A. Limaye, P. Pottinger, E. Whimbey, C. H. Goss, M. R. Tonelli, G. A. Cangelosi, M. A. Dirac, K. N. Olivier, B. A. Brown-Elliott, S. McNulty and R. J. Wallace, Jr. (2012). "Respiratory outbreak of Mycobacterium abscessus subspecies massiliense in a lung transplant and cystic fibrosis center." American Journal of Respiratory and Critical Care Medicine **185**(2): 231-232.

Akhter, Y., S. Yellaboina, A. Farhana, A. Ranjan, N. Ahmed and S. E. Hasnain (2008). "Genome scale portrait of cAMP-receptor protein (CRP) regulons in mycobacteria points to their role in pathogenesis." Gene **407**(1-2): 148-158.

Appelgren, P., F. Farnebo, L. Dotevall, M. Studahl, B. Jonsson and B. Petrini (2008). "Late-onset posttraumatic skin and soft-tissue infections caused by rapid-growing mycobacteria in tsunami survivors." Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **47**(2): e11-16.

Bange, F. C., B. A. Brown, C. Smaczny, R. J. Wallace Jr and E. C. Bottger (2001). "Lack of transmission of mycobacterium abscessus among patients with cystic fibrosis attending a single clinic." Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **32**(11): 1648-1650.

Bar-On, O., H. Mussaffi, M. Mei-Zahav, D. Prais, G. Steuer, P. Stafler, S. Hananya and H. Blau (2014). "Increasing nontuberculous mycobacteria infection in cystic fibrosis." Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society.

Bentley, S. D., I. Comas, J. M. Bryant, D. Walker, N. H. Smith, S. R. Harris, S. Thurston, S. Gagneux, J. Wood, M. Antonio, M. A. Quail, F. Gehre, R. A. Adegbola, J. Parkhill and B. C. de Jong (2012). "The genome of Mycobacterium africanum West African 2 reveals a lineage-specific locus and genome erosion common to the M. tuberculosis complex." PLoS neglected tropical diseases **6**(2): e1552.

Bicego, G. T., R. Nkambule, I. Peterson, J. Reed, D. Donnell, H. Ginindza, Y. T. Duong, H. Patel, N. Bock, N. Philip, C. Mao and J. Justman (2013). "Recent patterns in population-based HIV prevalence in Swaziland." PloS one **8**(10): e77101.

Blauwendraat, C., G. L. Dixon, J. C. Hartley, J. Foweraker and K. A. Harris (2012). "The use of a two-gene sequencing approach to accurately distinguish between the species within the Mycobacterium abscessus complex and Mycobacterium chelonae." European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology.

Bos, K. and J. Krause (in press). "Pre-Columbian Mycobacterial Genomes Reveal Seals as a Source of New World Human Tuberculosis." Nature.

Broda, A., H. Jebbari, K. Beaton, S. Mitchell and F. Drobniewski (2013). "Comparative drug resistance of Mycobacterium abscessus and M. chelonae isolates from patients with and without cystic fibrosis in the United Kingdom." Journal of Clinical Microbiology **51**(1): 217-223.

Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L. M., Pym, A. S., Samper, S., van Soolingen, D. and Cole, S. T (2002). "A new evolutionary scenario for the Mycobacterium tuberculosis complex,." PNAS **99**(6): 3684-3689.

Brown, T., V. Nikolayevskyy, P. Velji and F. Drobniewski (2010). "Associations between Mycobacterium tuberculosis strains and phenotypes." Emerging Infectious Diseases **16**(2): 272-280.

Cangelosi, G. A., R. J. Freeman, K. N. Lewis, D. Livingston-Rosanoff, K. S. Shah, S. J. Milan and S. V. Goldberg (2004). "Evaluation of a high-throughput repetitive-sequence-based PCR system for DNA fingerprinting of Mycobacterium tuberculosis and Mycobacterium avium complex strains." Journal of Clinical Microbiology **42**(6): 2685-2693.

Carver, T., S. R. Harris, M. Berriman, J. Parkhill and J. A. McQuillan (2011). "Artemis: An integrated platform for visualisation and analysis of high-throughput sequence-based experimental data." Bioinformatics.

Casali, N., V. Nikolayevskyy, Y. Balabanova, S. R. Harris, O. Ignatyeva, I. Kontsevaya, J. Corander, J. Bryant, J. Parkhill, S. Nejentsev, R. D. Horstmann, T. Brown and F. Drobniewski (2014). "Evolution and transmission of drug-resistant tuberculosis in a Russian population." Nature Genetics **46**(3): 279-286.

Casali, N., V. Nikolayevskyy, Y. Balabanova, O. Ignatyeva, I. Kontsevaya, S. R. Harris, S. D. Bentley, J. Parkhill, S. Nejentsev, S. E. Hoffner, R. D. Horstmann, T. Brown and F. Drobniewski (2012). "Microevolution of extensively drug-resistant tuberculosis in Russia." Genome Research **22**(4): 735-745.

Chan, J., M. Halachev, E. Yates, G. Smith and M. Pallen (2012). "Whole-genome sequence of the emerging pathogen Mycobacterium abscessus strain 47J26." Journal of Bacteriology **194**(2): 549.

Choi, G. E., Y. J. Cho, W. J. Koh, J. Chun, S. N. Cho and S. J. Shin (2012). "Draft genome sequence of Mycobacterium abscessus subsp. bolletii BD(T)." Journal of Bacteriology **194**(10): 2756-2757.

Choo, S. W., Y. L. Wong, M. L. Leong, H. Heydari, C. S. Ong, K. P. Ng and Y. F. Ngeow (2012). "Analysis of the genome of Mycobacterium abscessus strain M94 reveals an uncommon cluster of tRNAs." Journal of Bacteriology **194**(20): 5724.

Choo, S. W., Y. L. Wong, J. L. Tan, C. S. Ong, G. J. Wong, K. P. Ng and Y. F. Ngeow (2012). "Annotated genome sequence of Mycobacterium massiliense strain M154, belonging to the recently created taxon Mycobacterium abscessus subsp. bolletii comb. nov." Journal of Bacteriology **194**(17): 4778.

Choo, S. W., Y. L. Wong, A. M. Yusoff, M. L. Leong, G. J. Wong, C. S. Ong, K. P. Ng and Y. F. Ngeow (2012). "Genome sequence of the Mycobacterium abscessus strain M93." Journal of Bacteriology **194**(12): 3278.

Chou, M. P., A. C. Clements and R. M. Thomson (2014). "A spatial epidemiological analysis of nontuberculous mycobacterial infections in Queensland, Australia." BMC infectious diseases **14**(1): 279.

Chung, J. C., J. Becq, L. Fraser, O. Schulz-Trieglaff, N. J. Bond, J. Foweraker, K. D. Bruce, G. P. Smith and M. Welch (2012). "Genomic variation among contemporary Pseudomonas aeruginosa isolates from chronically infected cystic fibrosis patients." Journal of Bacteriology **194**(18): 4857-4866.

Clark, T. G., K. Mallard, F. Coll, M. Preston, S. Assefa, D. Harris, S. Ogwang, F. Mumbowa, B. Kirenga, D. M. O'Sullivan, A. Okwera, K. D. Eisenach, M. Joloba, S. D. Bentley, J. J. Ellner, J. Parkhill, E. C. Jones-Lopez and R. McNerney (2013). "Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing." PloS one **8**(12): e83012.

Cock, P. J., C. J. Fields, N. Goto, M. L. Heuer and P. M. Rice (2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." Nucleic Acids Research **38**(6): 1767-1771.

Cohen, T., P. D. van Helden, D. Wilson, C. Colijn, M. M. McLaughlin, I. Abubakar and R. M. Warren (2012). "Mixed-strain mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control." Clinical Microbiology Reviews **25**(4): 708-719.

Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead and B. G. Barrell (1998). "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence." Nature **393**(6685): 537-544.

Colijn, C., T. Cohen and M. Murray (2009). "Latent coinfection and the maintenance of strain diversity." Bulletin of Mathematical Biology **71**(1): 247-263.

Collins, F. S., M. Morgan and A. Patrinos (2003). "The Human Genome Project: lessons from large-scale biology." Science **300**(5617): 286-290.

Comas, I., S. Borrell, A. Roetzer, G. Rose, B. Malla, M. Kato-Maeda, J. Galagan, S. Niemann and S. Gagneux (2011). "Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes." Nature genetics **44**(1): 106-110.

Comas, I., M. Coscolla, T. Luo, S. Borrell, K. E. Holt, M. Kato-Maeda, J. Parkhill, B. Malla, S. Berg, G. Thwaites, D. Yeboah-Manu, G. Bothamley, J. Mei, L. Wei, S. Bentley, S. R. Harris, S. Niemann, R. Diel, A. Aseffa, Q. Gao, D. Young and S. Gagneux (2013). "Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans." Nature Genetics **45**(10): 1176-1182.

Comas, I. and S. Gagneux (2009). "The past and future of tuberculosis research." PLoS pathogens **5**(10): e1000600.

Copin, R., M. Coscolla, S. N. Seiffert, G. Bothamley, J. Sutherland, G. Mbayo, S. Gagneux and J. D. Ernst (2014). "Sequence diversity in the pe_pgrs genes of Mycobacterium tuberculosis is independent of human T cell recognition." mBio **5**(1): e00960-00913.

Cotruvo, J. A., Jr. and J. Stubbe (2008). "NrdI, a flavodoxin involved in maintenance of the diferric-tyrosyl radical cofactor in Escherichia coli class Ib ribonucleotide reductase." Proceedings of the National Academy of Sciences of the United States of America **105**(38): 14383-14388.

Croucher, N. J., S. R. Harris, C. Fraser, M. A. Quail, J. Burton, M. van der Linden, L. McGee, A. von Gottberg, J. H. Song, K. S. Ko, B. Pichon, S. Baker, C. M. Parry, L. M. Lambertsen, D. Shahinas, D. R. Pillai, T. J. Mitchell, G. Dougan, A. Tomasz, K. P. Klugman, J. Parkhill, W. P. Hanage and S. D. Bentley (2011). "Rapid pneumococcal evolution in response to clinical interventions." Science **331**(6016): 430-434.

Daniel, T. M. (2006). "The history of tuberculosis." Respiratory Medicine **100**(11): 1862-1870.

Davidson, R. M., N. A. Hasan, V. C. de Moura, R. S. Duarte, M. Jackson and M. Strong (2013). "Phylogenomics of Brazilian epidemic isolates of Mycobacterium abscessus subsp. bolletii reveals relationships of global outbreak strains." Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases **20**: 292-297.

de Beer, J. L., K. Kremer, C. Kodmon, P. Supply and D. van Soolingen (2012). "First worldwide proficiency study on variable-number tandem-repeat typing of Mycobacterium tuberculosis complex strains." Journal of Clinical Microbiology **50**(3): 662-669.

Delogu, G. and M. J. Brennan (2001). "Comparative immune response to PE and PE_PGRS antigens of Mycobacterium tuberculosis." Infection and Immunity **69**(9): 5606-5611.

Depledge, D. P., A. L. Palser, S. J. Watson, I. Y. Lai, E. R. Gray, P. Grant, R. K. Kanda, E. Leproust, P. Kellam and J. Breuer (2011). "Specific capture and whole-genome sequencing of viruses from clinical samples." PloS one **6**(11): e27805.

Devulder, G., M. Perouse de Montclos and J. P. Flandrois (2005). "A multigene approach to phylogenetic analysis using the genus Mycobacterium as a model." International journal of systematic and evolutionary microbiology **55**(Pt 1): 293-302.

Diacon, A. H., J. S. Maritz, A. Venter, P. D. van Helden, R. Dawson and P. R. Donald (2012). "Time to liquid culture positivity can substitute for colony counting on agar plates in early bactericidal activity studies of antituberculosis agents." Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **18**(7): 711-717.

Didelot, X., R. Bowden, D. J. Wilson, T. E. Peto and D. W. Crook (2012). "Transforming clinical microbiology with bacterial genome sequencing." Nature reviews. Genetics **13**(9): 601-612.

Drummond, A. J. and A. Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees." BMC evolutionary biology **7**: 214.

Dye, C. and B. G. Williams (2010). "The population dynamics and control of tuberculosis." Science **328**(5980): 856-861.

Dytoc, M. T., L. Honish, C. Shandro, P. T. Ting, L. Chui, L. Fiorillo, J. Robinson, A. Fanning, G. Predy and R. P. Rennie (2005). "Clinical, microbiological, and epidemiological findings of an outbreak of Mycobacterium abscessus hand-and-foot disease." Diagnostic Microbiology and Infectious Disease **53**(1): 39-45.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic acids research **32**(5): 1792-1797.

Elborn, J. S. (2013). "Current Approaches to the Management of Infection in Cystic Fibrosis." Current Pediatrics Reports **1**(3): 141-148.

Eyre, D. W., T. Golubchik, N. C. Gordon, R. Bowden, P. Piazza, E. M. Batty, C. L. Ip, D. J. Wilson, X. Didelot, L. O'Connor, R. Lay, D. Buck, A. M. Kearns, A. Shaw, J. Paul, M. H. Wilcox, P. J. Donnelly, T. E. Peto, A. S. Walker and D. W. Crook (2012). "A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance." BMJ open **2**(3).

Falkinham, J. O., 3rd (2002). "Nontuberculous mycobacteria in the environment." Clinics in Chest Medicine **23**(3): 529-551.

Falkinham, J. O., 3rd (2003). "Mycobacterial aerosols and respiratory disease." Emerging Infectious Diseases **9**(7): 763-767.

Farhat, M. R., B. J. Shapiro, K. J. Kieser, R. Sultana, K. R. Jacobson, T. C. Victor, R. M. Warren, E. M. Streicher, A. Calver, A. Sloutsky, D. Kaur, J. E. Posey, B. Plikaytis, M. R. Oggioni, J. L. Gardy, J. C. Johnston, M. Rodrigues, P. K. Tang, M. Kato-Maeda, M. L. Borowsky, B. Muddukrishna, B. N. Kreiswirth, N. Kurepina, J. Galagan, S. Gagneux, B. Birren, E. J. Rubin, E. S. Lander, P. C. Sabeti and M. Murray (2013). "Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis." Nature Genetics **45**(10): 1183-1189.

Farris (1970). "Methods for computing Wagner Trees." Syst. Zool., **19**: 83–92.

Fine, P. E. (1995). "Variation in protection by BCG: implications of and for heterologous immunity." Lancet **346**(8986): 1339-1345.

Fischer, A., I. Vazquez-Garcia, C. J. Illingworth and V. Mustonen (2014). "High-definition reconstruction of clonal composition in cancer." Cell reports **7**(5): 1740-1752.

Ford, C. B., P. L. Lin, M. R. Chase, R. R. Shah, O. Iartchouk, J. Galagan, N. Mohaideen, T. R. Ioerger, J. C. Sacchettini, M. Lipsitch, J. L. Flynn and S. M.

Fortune (2011). "Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection." Nature genetics **43**(5): 482-486.

Ford, C. B., R. R. Shah, M. K. Maeda, S. Gagneux, M. B. Murray, T. Cohen, J. C. Johnston, J. Gardy, M. Lipsitch and S. M. Fortune (2013). "Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis." Nature Genetics **45**(7): 784-790.

Gagneux, S., K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. C. Gutierrez, M. Hilty, P. C. Hopewell and P. M. Small (2006). "Variable host-pathogen compatibility in Mycobacterium tuberculosis." Proceedings of the National Academy of Sciences of the United States of America **103**(8): 2869-2873.

Gangadharam, P. R. J. and P. A. Jenkins (1997). Mycobacteria I: Basic Aspects. New York, Springer.

Gardy, J. L., J. C. Johnston, S. J. Ho Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. Jones, F. S. Brinkman, R. C. Brunham and P. Tang (2011). "Whole-genome sequencing and social-network analysis of a tuberculosis outbreak." The New England journal of medicine **364**(8): 730-739.

Gengenbacher, M., T. Xu, P. Niyomrattanakit, G. Spraggon and T. Dick (2008). "Biochemical and structural characterization of the putative dihydropteroate synthase ortholog Rv1207 of Mycobacterium tuberculosis." FEMS Microbiology Letters **287**(1): 128-135.

Gey Van Pittius, N. C., J. Gamieldien, W. Hide, G. D. Brown, R. J. Siezen and A. D. Beyers (2001). "The ESAT-6 gene cluster of Mycobacterium tuberculosis and other high G+C Gram-positive bacteria." Genome biology **2**(10): RESEARCH0044.

Gey van Pittius, N. C., S. L. Sampson, H. Lee, Y. Kim, P. D. van Helden and R. M. Warren (2006). "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 evolutionary biology **6**: 95.

Gillespie, S. H., A. M. Crook, T. D. McHugh, C. M. Mendel, S. K. Meredith, S. R. Murray, F. Pappas, P. P. Phillips, A. J. Nunn and R. Consortium (2014). "Four-month moxifloxacin-based regimens for drug-sensitive tuberculosis." N Engl J Med **371**(17): 1577-1587.

Glynn, J. R., J. Murray, A. Bester, G. Nelson, S. Shearer and P. Sonnenberg (2010). "High rates of recurrence in HIV-infected and HIV-uninfected patients with tuberculosis." The Journal of infectious diseases **201**(5): 704-711.

Glynn, J. R., M. D. Yates, A. C. Crampin, B. M. Ngwira, F. D. Mwaungulu, G. F. Black, S. D. Chaguluka, D. T. Mwafulirwa, S. Floyd, C. Murphy, F. A. Drobniewski and P. E. Fine (2004). "DNA fingerprint changes in tuberculosis: reinfection, evolution, or laboratory error?" The Journal of infectious diseases **190**(6): 1158-1166.

Golubchik, T., E. M. Batty, R. R. Miller, H. Farr, B. C. Young, H. Larner-Svensson, R. Fung, H. Godwin, K. Knox, A. Votintseva, R. G. Everitt, T. Street, M. Cule, C. L. C. Ip, X. Didelot, T. E. A. Peto, R. M. Harding, D. J. Wilson, D. W. Crook and R. Bowden (2013). "Within-Host Evolution of Staphylococcus aureus during Asymptomatic Carriage." PloS one **8**(5).

Govan, J. R. and V. Deretic (1996). "Microbial pathogenesis in cystic fibrosis: mucoid Pseudomonas aeruginosa and Burkholderia cepacia." Microbiological Reviews **60**(3): 539-574.

Griffith, D. E. (2003). "Emergence of nontuberculous mycobacteria as pathogens in cystic fibrosis." American journal of respiratory and critical care medicine **167**(6): 810-812.

Hanekom, M., E. M. Streicher, D. Van de Berg, H. Cox, C. McDermid, M. Bosman, N. C. Gey van Pittius, T. C. Victor, M. Kidd, D. van Soolingen, P. D. van Helden and R. M. Warren (2013). "Population structure of mixed Mycobacterium tuberculosis infection is strain genotype and culture medium dependent." PloS one **8**(7): e70178.

Harris, S. R., E. J. Cartwright, M. E. Torok, M. T. Holden, N. M. Brown, A. L. Ogilvy-Stuart, M. J. Ellington, M. A. Quail, S. D. Bentley, J. Parkhill and S. J. Peacock (2013). "Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study." The Lancet infectious diseases **13**(2): 130-136.

Harris, S. R., E. J. Feil, M. T. Holden, M. A. Quail, E. K. Nickerson, N. Chantratita, S. Gardete, A. Tavares, N. Day, J. A. Lindsay, J. D. Edgeworth, H. de Lencastre, J. Parkhill, S. J. Peacock and S. D. Bentley (2010). "Evolution of MRSA during hospital transmission and intercontinental spread." Science **327**(5964): 469-474.

Helb, D., M. Jones, E. Story, C. Boehme, E. Wallace, K. Ho, J. Kop, M. R. Owens, R. Rodgers, P. Banada, H. Safi, R. Blakemore, N. T. Lan, E. C. Jones-Lopez, M. Levi, M. Burday, I. Ayakaka, R. D. Mugerwa, B. McMillan, E. Winn-Deen, L. Christel, P. Dailey, M. D. Perkins, D. H. Persing and D. Alland (2010). "Rapid detection of Mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology." Journal of Clinical Microbiology **48**(1): 229-237.

Hermon-Taylor, J. and F. El-Zaatari (2004). The Mycobacterium avium subspecies paratuberculosis problem and its relation to the causation of Crohn disease. London, IWA Publishing.

Hershberg, R., M. Lipatov, P. M. Small, H. Sheffer, S. Niemann, S. Homolka, J. C. Roach, K. Kremer, D. A. Petrov, M. W. Feldman and S. Gagneux (2008). "High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography." PLoS biology **6**(12): e311.

Hershkovitz, I., H. D. Donoghue, D. E. Minnikin, G. S. Besra, O. Y. Lee, A. M. Gernaey, E. Galili, V. Eshed, C. L. Greenblatt, E. Lemma, G. K. Bar-Gal and M. Spigelman (2008). "Detection and molecular characterization of 9,000-year-old Mycobacterium tuberculosis from a Neolithic settlement in the Eastern Mediterranean." PloS one **3**(10): e3426.

Heym, B., P. M. Alzari, N. Honore and S. T. Cole (1995). "Missense mutations in the catalase-peroxidase gene, katG, are associated with isoniazid resistance in Mycobacterium tuberculosis." Molecular microbiology **15**(2): 235-245.

Ho, S. Y., B. Shapiro, M. J. Phillips, A. Cooper and A. J. Drummond (2007). "Evidence for time dependency of molecular rate estimates." Systematic biology **56**(3): 515-522.

Hoefsloot, W., J. van Ingen, C. Andrejak, K. Angeby, R. Bauriaud, P. Bemer, N. Beylis, M. J. Boeree, J. Cacho, V. Chihota, E. Chimara, G. Churchyard, R. Cias, R. Daza, C. L. Daley, P. N. Dekhuijzen, D. Domingo, F. Drobniewski, J. Esteban, M. Fauville-Dufaux, D. B. Folkvardsen, N. Gibbons, E. Gomez-Mampaso, R. Gonzalez, H. Hoffmann, P. R. Hsueh, A. Indra, T. Jagielski, F. Jamieson, M. Jankovic, E. Jong, J. Keane, W. J. Koh, B. Lange, S. Leao, R. Macedo, T. Mannsaker, T. K. Marras, J. Maugein, H. J. Milburn, T. Mlinko, N. Morcillo, K. Morimoto, D. Papaventsis, E. Palenque, M. Paez-Pena, C. Piersimoni, M. Polanova, N. Rastogi, E. Richter, M. J. Ruiz-Serrano, A. Silva, M. P. da Silva, H. Simsek, D. van Soolingen, N. Szabo, R. Thomson, T. Tortola Fernandez, E. Tortoli, S. E. Totten, G. Tyrrell, T. Vasankari, M. Villar, R. Walkiewicz, K. L. Winthrop and D. Wagner (2013). "The geographic diversity of nontuberculous mycobacteria isolated from pulmonary samples: an NTM-NET collaborative study." The European respiratory journal **42**(6): 1604-1613.

Institute, C. a. L. S. (2011). Susceptibility Testing of Mycobacteria, Nocardiae, and Other Aerobic Actinomycetes; Approved Standard - Second Edition.

Jiang, X., W. Zhang, F. Gao, Y. Huang, C. Lv and H. Wang (2006). "Comparison of the proteome of isoniazid-resistant and -susceptible strains of Mycobacterium tuberculosis." Microbial drug resistance **12**(4): 231-238.

Johnson, P. D., T. Stinear, P. L. Small, G. Pluschke, R. W. Merritt, F. Portaels, K. Huygen, J. A. Hayman and K. Asiedu (2005). "Buruli ulcer (M. ulcerans infection): new insights, new hope for disease control." PLoS medicine **2**(4): e108.

Kamerbeek, J., L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal and J. van Embden (1997).

"Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology." J Clin Microbiol **35**(4): 907-914.

Kamerbeek, J., L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal and J. van Embden (1997). "Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology." Journal of Clinical Microbiology **35**(4): 907-914.

Karboul, A., A. Mazza, N. C. Gey van Pittius, J. L. Ho, R. Brousseau and H. Mardassi (2008). "Frequent homologous recombination events in Mycobacterium tuberculosis PE/PPE multigene families: potential role in antigenic variability." Journal of bacteriology **190**(23): 7838-7846.

Katoh, K. and D. M. Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." Molecular Biology and Evolution **30**(4): 772-780.

Kennemann, L., X. Didelot, T. Aebischer, S. Kuhn, B. Drescher, M. Droege, R. Reinhardt, P. Correa, T. F. Meyer, C. Josenhans, D. Falush and S. Suerbaum (2011). "Helicobacter pylori genome evolution during human infection." Proceedings of the National Academy of Sciences of the United States of America **108**(12): 5033-5038.

Kim, K. H., D. R. An, J. Song, J. Y. Yoon, H. S. Kim, H. J. Yoon, H. N. Im, J. Kim, J. Kim do, S. J. Lee, H. M. Lee, H. J. Kim, E. K. Jo, J. Y. Lee and S. W. Suh (2012). "Mycobacterium tuberculosis Eis protein initiates suppression of host immune responses by acetylation of DUSP16/MKP-7." Proceedings of the National Academy of Sciences of the United States of America **109**(20): 7729-7734.

Klopper, M., R. M. Warren, C. Hayes, N. C. Gey van Pittius, E. M. Streicher, B. Muller, F. A. Sirgel, M. Chabula-Nxiweni, E. Hoosain, G. Coetzee, P. David van Helden, T. C. Victor and A. P. Trollip (2013). "Emergence and spread of extensively and totally drug-resistant tuberculosis, South Africa." Emerging Infectious Diseases **19**(3): 449-455.

Knechel, N. A. (2009). "Tuberculosis: pathophysiology, clinical features, and diagnosis." Critical Care Nurse **29**(2): 34-43; quiz 44.

Koh, S. J., T. Song, Y. A. Kang, J. W. Choi, K. J. Chang, C. S. Chu, J. G. Jeong, J. Y. Lee, M. K. Song, H. Y. Sung, Y. H. Kang and J. J. Yim (2010). "An outbreak of skin and soft tissue infection caused by Mycobacterium abscessus following acupuncture." Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **16**(7): 895-901.

Konstantinidis, K. T. and J. M. Tiedje (2007). "Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead." Current Opinion in Microbiology **10**(5): 504-509.

Koser, C. U., M. J. Ellington, E. J. Cartwright, S. H. Gillespie, N. M. Brown, M. Farrington, M. T. Holden, G. Dougan, S. D. Bentley, J. Parkhill and S. J. Peacock (2012). "Routine use of microbial whole genome sequencing in diagnostic and public health microbiology." PLoS pathogens **8**(8): e1002824.

Koser, C. U., M. T. Holden, M. J. Ellington, E. J. Cartwright, N. M. Brown, A. L. Ogilvy-Stuart, L. Y. Hsu, C. Chewapreecha, N. J. Croucher, S. R. Harris, M. Sanders, M. C. Enright, G. Dougan, S. D. Bentley, J. Parkhill, L. J. Fraser, J. R. Betley, O. B. Schulz-Trieglaff, G. P. Smith and S. J. Peacock (2012). "Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak." The New England journal of medicine **366**(24): 2267-2275.

Kremer, K., M. J. van-der-Werf, B. K. Au, D. D. Anh, K. M. Kam, H. R. van-Doorn, M. W. Borgdorff and D. van-Soolingen (2009). "Vaccine-induced immunity circumvented by typical Mycobacterium tuberculosis Beijing strains." Emerging Infectious Diseases **15**(2): 335-339.

Kreutzfeldt, K. M., P. R. McAdam, P. Claxton, A. Holmes, A. L. Seagar, I. F. Laurenson and J. R. Fitzgerald (2013). "Molecular longitudinal tracking of Mycobacterium abscessus spp. during chronic infection of the human lung." PloS one **8**(5): e63237.

Krishnan, N., W. Malaga, P. Constant, M. Caws, T. H. Tran, J. Salmons, T. N. Nguyen, D. B. Nguyen, M. Daffe, D. B. Young, B. D. Robertson, C. Guilhot and G. E. Thwaites (2011). "Mycobacterium tuberculosis lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles." PloS one **6**(9): e23870.

Kusunoki, S. and T. Ezaki (1992). "Proposal of Mycobacterium peregrinum sp. nov., nom. rev., and elevation of Mycobacterium chelonae subsp. abscessus (Kubica et al.) to species status: Mycobacterium abscessus comb. nov." International journal of systematic bacteriology **42**(2): 240-245.

Lai, C. C., C. K. Tan, C. H. Chou, H. L. Hsu, C. H. Liao, Y. T. Huang, P. C. Yang, K. T. Luh and P. R. Hsueh (2010). "Increasing incidence of nontuberculous mycobacteria, Taiwan, 2000-2008." Emerging Infectious Diseases **16**(2): 294-296.

Leao, S. C., E. Tortoli, J. P. Euzeby and M. J. Garcia (2011). "Proposal that Mycobacterium massiliense and Mycobacterium bolletii be united and reclassified as Mycobacterium abscessus subsp. bolletii comb. nov., designation of Mycobacterium abscessus subsp. abscessus subsp. nov. and emended description of Mycobacterium abscessus." International journal of systematic and evolutionary microbiology **61**(Pt 9): 2311-2313.

Leao, S. C., C. Viana-Niero, C. K. Matsumoto, K. V. Lima, M. L. Lopes, M. Palaci, D. J. Hadad, S. Vinhas, R. S. Duarte, M. C. Lourenco, A. Kipnis, Z. C. das Neves, B. M. Gabardo, M. O. Ribeiro, L. Baethgen, D. B. de Assis, G. Madalosso, E. Chimara and M. P. Dalcolmo (2010). "Epidemic of surgical-site infections by a single clone of rapidly growing mycobacteria in Brazil." Future microbiology **5**(6): 971-980.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Liao, X. and R. E. Hancock (1997). "Susceptibility to beta-lactam antibiotics of Pseudomonas aeruginosa overproducing penicillin-binding protein 3." Antimicrob Agents Chemother **41**(5): 1158-1161.

Lieberman, T. D., K. B. Flett, I. Yelin, T. R. Martin, A. J. McAdam, G. P. Priebe and R. Kishony (2014). "Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures." Nature Genetics **46**(1): 82-87.

Lieberman, T. D., J. B. Michel, M. Aingaran, G. Potter-Bynoe, D. Roux, M. R. Davis, Jr., D. Skurnik, N. Leiby, J. J. LiPuma, J. B. Goldberg, A. J. McAdam, G. P. Priebe and R. Kishony (2011). "Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes." Nature genetics **43**(12): 1275-1280.

Lin, P. L., C. B. Ford, M. T. Coleman, A. J. Myers, R. Gawande, T. Ioerger, J. Sacchettini, S. M. Fortune and J. L. Flynn (2014). "Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing." Nature Medicine **20**(1): 75-79.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law (2012). "Comparison of next-generation sequencing systems." Journal of biomedicine & biotechnology **2012**: 251364.

Liu, X., M. M. Gutacker, J. M. Musser and Y. X. Fu (2006). "Evidence for recombination in Mycobacterium tuberculosis." Journal of Bacteriology **188**(23): 8169-8177.

Lopez, B., D. Aguilar, H. Orozco, M. Burger, C. Espitia, V. Ritacco, L. Barrera, K. Kremer, R. Hernandez-Pando, K. Huygen and D. van Soolingen (2003). "A marked difference in pathogenesis and immune response induced by different Mycobacterium tuberculosis genotypes." Clinical and Experimental Immunology **133**(1): 30-37.

LPSN. (2014). "List of prokaryotic names with standing in nomenclature: Genus Mycobacterium." Retrieved January, 2014, from http://www.bacterio.net/mycobacterium.html.

Macheras, E., J. Konjek, A. L. Roux, J. M. Thiberge, S. Bastian, S. C. Leao, M. Palaci, V. Sivadon-Tardy, C. Gutierrez, E. Richter, S. Rusch-Gerdes, G. E. Pfyffer, T. Bodmer, V. Jarlier, E. Cambau, S. Brisse, V. Caro, N. Rastogi, J. L. Gaillard and B.

Heym (2013). "Multilocus sequence typing scheme for the Mycobacterium abscessus complex." Research in Microbiology.

Macheras, E., A. L. Roux, S. Bastian, S. C. Leao, M. Palaci, V. Sivadon-Tardy, C. Gutierrez, E. Richter, S. Rusch-Gerdes, G. Pfyffer, T. Bodmer, E. Cambau, J. L. Gaillard and B. Heym (2011). "Multilocus sequence analysis and rpoB sequencing of Mycobacterium abscessus (sensu lato) strains." Journal of Clinical Microbiology **49**(2): 491-499.

Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman and B. G. Spratt (1998). "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms." Proceedings of the National Academy of Sciences of the United States of America **95**(6): 3140-3145.

Martin, A., M. Herranz, Y. Navarro, S. Lasarte, M. J. Ruiz Serrano, E. Bouza and D. Garcia de Viedma (2011). "Evaluation of the inaccurate assignment of mixed infections by Mycobacterium tuberculosis as exogenous reinfection and analysis of the potential role of bacterial factors in reinfection." Journal of Clinical Microbiology **49**(4): 1331-1338.

Matsumoto, C. K., P. J. Bispo, K. Santin, C. L. Nogueira and S. C. Leao (2014). "Demonstration of plasmid-mediated drug resistance in Mycobacterium abscessus." Journal of Clinical Microbiology **52**(5): 1727-1729.

Matsumoto, C. K., E. Chimara, S. Bombarda, R. S. Duarte and S. C. Leao (2011). "Diversity of pulsed-field gel electrophoresis patterns of Mycobacterium abscessus type 2 clinical isolates." Journal of Clinical Microbiology **49**(1): 62-68.

Maus, C. E., B. B. Plikaytis and T. M. Shinnick (2005). "Molecular analysis of cross-resistance to capreomycin, kanamycin, amikacin, and viomycin in Mycobacterium tuberculosis." Antimicrobial agents and chemotherapy **49**(8): 3192-3197.

McEvoy, C. R., R. Cloete, B. Muller, A. C. Schurch, P. D. van Helden, S. Gagneux, R. M. Warren and N. C. Gey van Pittius (2012). "Comparative analysis of

Mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints." PloS one **7**(4): e30593.

McEvoy, C. R., P. D. van Helden, R. M. Warren and N. C. Gey van Pittius (2009). "Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic Mycobacterium tuberculosis PPE38 gene region." BMC evolutionary biology **9**: 237.

Medjahed, H., J. L. Gaillard and J. M. Reyrat (2010). "Mycobacterium abscessus: a new player in the mycobacterial field." Trends in microbiology **18**(3): 117-123.

Mestre, O., T. Luo, T. Dos Vultos, K. Kremer, A. Murray, A. Namouchi, C. Jackson, J. Rauzier, P. Bifani, R. Warren, V. Rasolofo, J. Mei, Q. Gao and B. Gicquel (2011). "Phylogeny of Mycobacterium tuberculosis Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair." PloS one **6**(1): e16020.

Microsoft Corporation (2011). Microsoft Excel for Mac 2011 version 14.3.9.

Morgulis, A., E. M. Gertz, A. A. Schaffer and R. Agarwala (2006). "A fast and symmetric DUST implementation to mask low-complexity DNA sequences." Journal of computational biology : a journal of computational molecular cell biology **13**(5): 1028-1040.

Movahedzadeh, F., D. A. Smith, R. A. Norman, P. Dinadayala, J. Murray-Rust, D. G. Russell, S. L. Kendall, S. C. Rison, M. S. McAlister, G. J. Bancroft, N. Q. McDonald, M. Daffe, Y. Av-Gay and N. G. Stoker (2004). "The Mycobacterium tuberculosis ino1 gene is essential for growth and virulence." Molecular microbiology **51**(4): 1003-1014.

Mutreja, A., D. W. Kim, N. R. Thomson, T. R. Connor, J. H. Lee, S. Kariuki, N. J. Croucher, S. Y. Choi, S. R. Harris, M. Lebens, S. K. Niyogi, E. J. Kim, T. Ramamurthy, J. Chun, J. L. Wood, J. D. Clemens, C. Czerkinsky, G. B. Nair, J. Holmgren, J. Parkhill and G. Dougan (2011). "Evidence for several waves of global transmission in the seventh cholera pandemic." Nature **477**(7365): 462-465.

Nakamura, K., T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara and S. Kanaya (2011). "Sequence-specific error profile of Illumina sequencers." Nucleic Acids Research **39**(13): e90.

Nakanaga, K., Y. Hoshino, Y. Era, K. Matsumoto, Y. Kanazawa, A. Tomita, M. Furuta, M. Washizu, M. Makino and N. Ishii (2011). "Multiple cases of cutaneous Mycobacterium massiliense infection in a "hot spa" in Japan." Journal of Clinical Microbiology **49**(2): 613-617.

Namouchi, A., X. Didelot, U. Schock, B. Gicquel and E. P. Rocha (2012). "After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection." Genome Research **22**(4): 721-734.

Narayanan, S., S. Swaminathan, P. Supply, S. Shanmugam, G. Narendran, L. Hari, R. Ramachandran, C. Locht, M. S. Jawahar and P. R. Narayanan (2010). "Impact of HIV Infection on the Recurrence of Tuberculosis in South India." Journal of Infectious Diseases **201**(5): 691-703.

Nash, K. A., B. A. Brown-Elliott and R. J. Wallace, Jr. (2009). "A novel gene, erm(41), confers inducible macrolide resistance to clinical isolates of Mycobacterium abscessus but is absent from Mycobacterium chelonae." Antimicrobial Agents and Chemotherapy **53**(4): 1367-1376.

National Jewish Health. (2014). "Nontuberculous Mycobacteria (NTM): Causes." from http://www.nationaljewish.org/healthinfo/conditions/ntm/causes.

Nessar, R., E. Cambau, J. M. Reyrat, A. Murray and B. Gicquel (2012). "Mycobacterium abscessus: a new antibiotic nightmare." The Journal of antimicrobial chemotherapy **67**(4): 810-818.

Ngeow, Y. F., W. Y. Wee, Y. L. Wong, J. L. Tan, C. S. Ongi, K. P. Ng and S. W. Choo (2012). "Genomic analysis of Mycobacterium abscessus strain M139, which has

an ambiguous subspecies taxonomic position." Journal of Bacteriology **194**(21): 6002-6003.

Ngeow, Y. F., Y. L. Wong, N. Lokanathan, G. J. Wong, C. S. Ong, K. P. Ng and S. W. Choo (2012). "Genomic Analysis of Mycobacterium massiliense strain M115, an isolate from human sputum." Journal of Bacteriology **194**(17): 4786.

Ngeow, Y. F., Y. L. Wong, J. L. Tan, C. S. Ong, K. P. Ng and S. W. Choo (2012). "Genome Sequence of Mycobacterium abscessus Strain M152." Journal of Bacteriology **194**(23): 6662.

Niemann, S., C. U. Koser, S. Gagneux, C. Plinke, S. Homolka, H. Bignell, R. J. Carter, R. K. Cheetham, A. Cox, N. A. Gormley, P. Kokko-Gonzales, L. J. Murray, R. Rigatti, V. P. Smith, F. P. M. Arends, H. S. Cox, G. Smith and J. A. C. Archer (2009). "Genomic Diversity among Drug Sensitive and Multidrug Resistant Isolates of *Mycobacterium tuberculosis* with Identical DNA Fingerprints." PLoS ONE **4**(10): e7407.

O'Sullivan, B. P. and C. M. Sassetti (2013). "Infection control in cystic fibrosis: share and share alike." Lancet **381**(9877): 1517-1519.

Ochman, H., S. Elwyn and N. A. Moran (1999). "Calibrating bacterial evolution." Proceedings of the National Academy of Sciences of the United States of America **96**(22): 12638-12643.

Oliver, A., R. Canton, P. Campo, F. Baquero and J. Blazquez (2000). "High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection." Science **288**(5469): 1251-1254.

Ordway, D., M. Henao-Tamayo, E. Smith, C. Shanley, M. Harton, J. Troudt, X. Bai, R. J. Basaraba, I. M. Orme and E. D. Chan (2008). "Animal model of Mycobacterium abscessus lung infection." Journal of Leukocyte Biology **83**(6): 1502-1511.

Perez-Lago, L., I. Comas, Y. Navarro, F. Gonzalez-Candelas, M. Herranz, E. Bouza and D. Garcia-de-Viedma (2014). "Whole genome sequencing analysis of intrapatient

microevolution in Mycobacterium tuberculosis: potential impact on the inference of tuberculosis transmission." The Journal of infectious diseases **209**(1): 98-108.

Pietersen, E., E. Ignatius, E. M. Streicher, B. Mastrapa, X. Padanilam, A. Pooran, M. Badri, M. Lesosky, P. van Helden, F. A. Sirgel, R. Warren and K. Dheda (2014). "Long-term outcomes of patients with extensively drug-resistant tuberculosis in South Africa: a cohort study." The Lancet.

Pilcher, C. D., J. K. Wong and S. K. Pillai (2008). "Inferring HIV transmission dynamics from phylogenetic sequence relationships." PLoS medicine **5**(3): e69.

Ponstingl, H. (2011). "SMALT v0.5.8." Retrieved 5/1, 2012, from http://www.sanger.ac.uk/resources/software/smalt/.

Prammananan, T., P. Sander, B. A. Brown, K. Frischkorn, G. O. Onyi, Y. Zhang, E. C. Bottger and R. J. Wallace, Jr. (1998). "A single 16S ribosomal RNA substitution is responsible for resistance to amikacin and other 2-deoxystreptamine aminoglycosides in Mycobacterium abscessus and Mycobacterium chelonae." The Journal of infectious diseases **177**(6): 1573-1581.

Prevots, D. R., P. A. Shaw, D. Strickland, L. A. Jackson, M. A. Raebel, M. A. Blosky, R. Montes de Oca, Y. R. Shea, A. E. Seitz, S. M. Holland and K. N. Olivier (2010). "Nontuberculous mycobacterial lung disease prevalence at four integrated health care delivery systems." American Journal of Respiratory and Critical Care Medicine **182**(7): 970-976.

Public Health England (2014). Collaborative Tuberculosis Stategy for England, 2014 to 2019. London, Public Health England.

Quail, M. A., T. D. Otto, Y. Gu, S. R. Harris, T. F. Skelly, J. A. McQuillan, H. P. Swerdlow and S. O. Oyola (2011). "Optimal enzymes for amplifying sequencing libraries." Nature methods **9**(1): 10-11.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow and Y. Gu (2012). "A tale of three next generation

sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." BMC genomics **13**: 341.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing.

Rachman, H., M. Strong, T. Ulrichs, L. Grode, J. Schuchhardt, H. Mollenkopf, G. A. Kosmiadi, D. Eisenberg and S. H. Kaufmann (2006). "Unique transcriptome signature of Mycobacterium tuberculosis in pulmonary tuberculosis." Infection and Immunity **74**(2): 1233-1242.

Raiol, T., G. M. Ribeiro, A. Q. Maranhao, A. L. Bocca, I. Silva-Pereira, A. P. Junqueira-Kipnis, M. Brigido Mde and A. Kipnis (2012). "Complete genome sequence of Mycobacterium massiliense." Journal of Bacteriology **194**(19): 5455.

Ramakrishnan, L., N. A. Federspiel and S. Falkow (2000). "Granuloma-specific expression of Mycobacterium virulence proteins from the glycine-rich PE-PGRS family." Science **288**(5470): 1436-1439.

Ramaswamy, S. and J. M. Musser (1998). "Molecular genetic basis of antimicrobial agent resistance in Mycobacterium tuberculosis: 1998 update." Tubercle and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease **79**(1): 3-29.

Rambaut, A. (2007). "Path-O-Gen." Retrieved 12/01, 2012, from http://tree.bio.ed.ac.uk/software/pathogen/.

Renna, M., C. Schaffner, K. Brown, S. Shang, M. H. Tamayo, K. Hegyi, N. J. Grimsey, D. Cusens, S. Coulter, J. Cooper, A. R. Bowden, S. M. Newton, B. Kampmann, J. Helm, A. Jones, C. S. Haworth, R. J. Basaraba, M. A. DeGroote, D. J. Ordway, D. C. Rubinsztein and R. A. Floto (2011). "Azithromycin blocks autophagy and may predispose cystic fibrosis patients to mycobacterial infection." The Journal of clinical investigation **121**(9): 3554-3563.

Richardson, M., G. D. van der Spuy, S. L. Sampson, N. Beyers, P. D. van Helden and R. M. Warren (2004). "Stability of polymorphic GC-rich repeat sequence-containing regions of Mycobacterium tuberculosis." Journal of Clinical Microbiology **42**(3): 1302-1304.

Riley, R. L. (1957). "Aerial dissemination of pulmonary tuberculosis." American review of tuberculosis **76**(6): 931-941.

Riley, R. L., C. C. Mills, W. Nyka, N. Weinstock, P. B. Storey, L. U. Sultan, M. C. Riley and W. F. Wells (1995). "Aerial dissemination of pulmonary tuberculosis. A two-year study of contagion in a tuberculosis ward. 1959." American Journal of Epidemiology **142**(1): 3-14.

Rinder, H. (2001). "Hetero-resistance: an under-recognised confounder in diagnosis and therapy?" Journal of Medical Microbiology **50**(12): 1018-1020.

Ripoll, F., C. Deshayes, S. Pasek, F. Laval, J. L. Beretti, F. Biet, J. L. Risler, M. Daffe, G. Etienne, J. L. Gaillard and J. M. Reyrat (2007). "Genomics of glycopeptidolipid biosynthesis in Mycobacterium abscessus and M. chelonae." BMC genomics **8**: 114.

Ripoll, F., S. Pasek, C. Schenowitz, C. Dossat, V. Barbe, M. Rottman, E. Macheras, B. Heym, J. L. Herrmann, M. Daffe, R. Brosch, J. L. Risler and J. L. Gaillard (2009). "Non mycobacterial virulence genes in the genome of the emerging pathogen Mycobacterium abscessus." PloS one **4**(6): e5660.

Roetzer, A., R. Diel, T. A. Kohl, C. Ruckert, U. Nubel, J. Blom, T. Wirth, S. Jaenicke, S. Schuback, S. Rusch-Gerdes, P. Supply, J. Kalinowski and S. Niemann (2013). "Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study." PLoS medicine **10**(2): e1001387.

Rook, G. A., K. Dheda and A. Zumla (2005). "Immune responses to tuberculosis in developing countries: implications for new vaccines." Nature reviews. Immunology **5**(8): 661-667.

Ruddy, M., T. D. McHugh, J. W. Dale, D. Banerjee, H. Maguire, P. Wilson, F. Drobniewski, P. Butcher and S. H. Gillespie (2002). "Estimation of the rate of unrecognized cross-contamination with mycobacterium tuberculosis in London microbiology laboratories." Journal of Clinical Microbiology **40**(11): 4100-4104.

Safi, H., S. Lingaraju, A. Amin, S. Kim, M. Jones, M. Holmes, M. McNeil, S. N. Peterson, D. Chatterjee, R. Fleischmann and D. Alland (2013). "Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-beta-D-arabinose biosynthetic and utilization pathway genes." Nature Genetics **45**(10): 1190-1197.

Salyers, A. A. (1995). Bacterial Pathogenesis: A Molecular Approach. Washington DC, American Society for Microbiology.

Sampson, S. L. (2011). "Mycobacterial PE/PPE proteins at the host-pathogen interface." Clinical & developmental immunology **2011**: 497203.

Sani, M., E. N. Houben, J. Geurtsen, J. Pierson, K. de Punder, M. van Zon, B. Wever, S. R. Piersma, C. R. Jimenez, M. Daffe, B. J. Appelmelk, W. Bitter, N. van der Wel and P. J. Peters (2010). "Direct visualization by cryo-EM of the mycobacterial capsular layer: a labile structure containing ESX-1-secreted proteins." PLoS pathogens **6**(3): e1000794.

Scaduto, D. I., J. M. Brown, W. C. Haaland, D. J. Zwickl, D. M. Hillis and M. L. Metzker (2010). "Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences." Proceedings of the National Academy of Sciences of the United States of America **107**(50): 21242-21247.

Schurch, A. C., K. Kremer, R. M. Warren, N. V. Hung, Y. Zhao, K. Wan, M. J. Boeree, R. J. Siezen, N. H. Smith and D. van Soolingen (2011). "Mutations in the regulatory network underlie the recent clonal expansion of a dominant subclone of the Mycobacterium tuberculosis Beijing genotype." Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases **11**(3): 587-597.

Sermet-Gaudelus, I., M. Le Bourgeois, C. Pierre-Audigier, C. Offredo, D. Guillemot, S. Halley, C. Akoua-Koffi, V. Vincent, V. Sivadon-Tardy, A. Ferroni, P. Berche, P. Scheinmann, G. Lenoir and J. L. Gaillard (2003). "Mycobacterium abscessus and children with cystic fibrosis." Emerg Infect Dis **9**(12): 1587-1591.

Shamputa, I. C., L. Jugheli, N. Sadradze, E. Willery, F. Portaels, P. Supply and L. Rigouts (2006). "Mixed infection and clonal representativeness of a single sputum sample in tuberculosis patients from a penitentiary hospital in Georgia." Respiratory research **7**: 99.

Shang, S., S. Gibbs, M. Henao-Tamayo, C. A. Shanley, G. McDonnell, R. S. Duarte, D. J. Ordway and M. Jackson (2011). "Increased virulence of an epidemic strain of Mycobacterium massiliense in mice." PloS one **6**(9): e24726.

Shin, D. M., B. Y. Jeon, H. M. Lee, H. S. Jin, J. M. Yuk, C. H. Song, S. H. Lee, Z. W. Lee, S. N. Cho, J. M. Kim, R. L. Friedman and E. K. Jo (2010). "Mycobacterium tuberculosis eis regulates autophagy, inflammation, and cell death through redox-dependent signaling." PLoS pathogens **6**(12): e1001230.

Smith, E. E., D. G. Buckley, Z. Wu, C. Saenphimmachak, L. R. Hoffman, D. A. D'Argenio, S. I. Miller, B. W. Ramsey, D. P. Speert, S. M. Moskowitz, J. L. Burns, R. Kaul and M. V. Olson (2006). "Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients." Proceedings of the National Academy of Sciences of the United States of America **103**(22): 8487-8492.

Smith, I. (2003). "Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence." Clinical Microbiology Reviews **16**(3): 463-496.

Sreevatsan, S., K. E. Stockbauer, X. Pan, B. N. Kreiswirth, S. L. Moghazeh, W. R. Jacobs, Jr., A. Telenti and J. M. Musser (1997). "Ethambutol resistance in Mycobacterium tuberculosis: critical role of embB mutations." Antimicrobial agents and chemotherapy **41**(8): 1677-1681.

Stamatakis, A. (2006). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics **22**(21): 2688-2690.

Stead, W. W. (1967). "Pathogenesis of a first episode of chronic pulmonary tuberculosis in man: recrudescence of residuals of the primary infection or exogenous reinfection?" The American review of respiratory disease **95**(5): 729-745.

Strong, M., M. R. Sawaya, S. Wang, M. Phillips, D. Cascio and D. Eisenberg (2006). "Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis." Proceedings of the National Academy of Sciences of the United States of America **103**(21): 8060-8065.

Stucki, D., B. Malla, S. Hostettler, T. Huna, J. Feldmann, D. Yeboah-Manu, S. Borrell, L. Fenner, I. Comas, M. Coscolla and S. Gagneux (2012). "Two new rapid SNP-typing methods for classifying Mycobacterium tuberculosis complex into the main phylogenetic lineages." PloS one **7**(7): e41253.

Sun, G., T. Luo, C. Yang, X. Dong, J. Li, Y. Zhu, H. Zheng, W. Tian, S. Wang, C. E. Barry, 3rd, J. Mei and Q. Gao (2012). "Dynamic Population Changes in Mycobacterium tuberculosis During Acquisition and Fixation of Drug Resistance in Patients." The Journal of infectious diseases.

Sun, G., T. Luo, C. Yang, X. Dong, J. Li, Y. Zhu, H. Zheng, W. Tian, S. Wang, C. E. Barry, 3rd, J. Mei and Q. Gao (2012). "Dynamic population changes in Mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients." The Journal of infectious diseases **206**(11): 1724-1733.

Supply, P., C. Allix, S. Lesjean, M. Cardoso-Oelemann, S. Rusch-Gerdes, E. Willery, E. Savine, P. de Haas, H. van Deutekom, S. Roring, P. Bifani, N. Kurepina, B. Kreiswirth, C. Sola, N. Rastogi, V. Vatin, M. C. Gutierrez, M. Fauville, S. Niemann, R. Skuce, K. Kremer, C. Locht and D. van Soolingen (2006). "Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of Mycobacterium tuberculosis." J. Clin. Microbiol. **44**(12): 4498-4510.

Supply, P., M. Marceau, S. Mangenot, D. Roche, C. Rouanet, V. Khanna, L. Majlessi, A. Criscuolo, J. Tap, A. Pawlik, L. Fiette, M. Orgeur, M. Fabre, C. Parmentier, W. Frigui, R. Simeone, E. C. Boritsch, A. S. Debrie, E. Willery, D. Walker, M. A. Quail, L. Ma, C. Bouchier, G. Salvignol, F. Sayes, A. Cascioferro, T. Seemann, V. Barbe, C. Locht, M. C. Gutierrez, C. Leclerc, S. D. Bentley, T. P. Stinear, S. Brisse, C. Medigue, J. Parkhill, S. Cruveiller and R. Brosch (2013). "Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis." Nature Genetics **45**(2): 172-179.

Takiff, H. E., L. Salazar, C. Guerrero, W. Philipp, W. M. Huang, B. Kreiswirth, S. T. Cole, W. R. Jacobs, Jr. and A. Telenti (1994). "Cloning and nucleotide sequence of Mycobacterium tuberculosis gyrA and gyrB genes and detection of quinolone resistance mutations." Antimicrobial agents and chemotherapy **38**(4): 773-780.

Talarico, S., M. D. Cave, C. F. Marrs, B. Foxman, L. Zhang and Z. Yang (2005). "Variation of the Mycobacterium tuberculosis PE_PGRS 33 gene among clinical isolates." Journal of Clinical Microbiology **43**(10): 4954-4960.

Talarico, S., L. Zhang, C. F. Marrs, B. Foxman, M. D. Cave, M. J. Brennan and Z. Yang (2008). "Mycobacterium tuberculosis PE_PGRS16 and PE_PGRS26 genetic polymorphism among clinical isolates." Tuberculosis **88**(4): 283-294.

Tettelin H, Davidson R.M, Agrawal S, Aitken M.L, Shallom S, Hasan N.A, Strong M and C. N. d. M. V (2014). "High-level relatedness among Mycobacterium abscessus subsp. massiliense strains from widely separated outbreaks." Emerg Infect Dis **[Ahead of print]**.

Tettelin, H., E. P. Sampaio, S. C. Daugherty, E. Hine, D. R. Riley, L. Sadzewicz, N. Sengamalay, K. Shefchek, Q. Su, L. J. Tallon, P. Conville, K. N. Olivier, S. M. Holland, C. M. Fraser and A. M. Zelazny (2012). "Genomic insights into the emerging human pathogen Mycobacterium massiliense." Journal of Bacteriology **194**(19): 5450.

Thomson, R. M. (2010). "Changing epidemiology of pulmonary nontuberculous mycobacteria infections." Emerging Infectious Diseases **16**(10): 1576-1583.

Torrens, J. K., P. Dawkins, S. P. Conway and E. Moya (1998). "Non-tuberculous mycobacteria in cystic fibrosis." Thorax **53**(3): 182-185.

Tripp, S. and M. Grueber (2011). Economic Impact of the Human Genome Project. B. M. Institute.

Tsenova, L., R. Harbacheuski, N. Sung, E. Ellison, D. Fallows and G. Kaplan (2007). "BCG vaccination confers poor protection against M. tuberculosis HN878-induced central nervous system disease." Vaccine **25**(28): 5126-5132.

Tsolaki, A. G., S. Gagneux, A. S. Pym, Y. O. Goguet de la Salmoniere, B. N. Kreiswirth, D. Van Soolingen and P. M. Small (2005). "Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis." Journal of Clinical Microbiology **43**(7): 3185-3191.

van Embden, J. D., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, T. M. Shinnick and et al. (1993). "Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology." Journal of Clinical Microbiology **31**(2): 406-409.

van Embden, J. D., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, T. M. Shinnick and et al. (1993). "Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology." J Clin Microbiol **31**(2): 406-409.

van Ingen, J., R. de Zwaan, R. P. Dekhuijzen, M. J. Boeree and D. van Soolingen (2009). "Clinical relevance of Mycobacterium chelonae-abscessus group isolation in 95 patients." The Journal of infection **59**(5): 324-331.

van Soolingen, D., L. Qian, P. E. de Haas, J. T. Douglas, H. Traore, F. Portaels, H. Z. Qing, D. Enkhsaikan, P. Nymadawa and J. D. van Embden (1995). "Predominance of a single genotype of Mycobacterium tuberculosis in countries of east Asia." Journal of Clinical Microbiology **33**(12): 3234-3238.

Vandamme, A. M. and O. G. Pybus (2013). "Viral phylogeny in court: the unusual case of the Valencian anesthetist." BMC biology **11**: 83.

Verver, S., R. M. Warren, N. Beyers, M. Richardson, G. D. van der Spuy, M. W. Borgdorff, D. A. Enarson, M. A. Behr and P. D. van Helden (2005). "Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis." American Journal of Respiratory and Critical Care Medicine **171**(12): 1430-1435.

Walker, T. M., C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, J. Parkhill, D. Harris, A. S. Walker, R. Bowden, P. Monk, E. G. Smith and T. E. Peto (2013). "Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study." The Lancet infectious diseases **13**(2): 137-146.

Walker, T. M., C. L. C. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, J. Parkhill, D. Harris, A. S. Walker, R. Bowden, P. Monk, E. G. Smith and T. E. A. Peto (2012). "Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study." The Lancet infectious diseases.

Walker, T. M., M. K. Lalor, A. Broda, L. S. Ortega, M. Morgan, L. Parker, S. Churchill, K. Bennett, T. Golubchik, A. P. Giess, C. Del Ojo Elias, K. J. Jeffery, I. C. Bowler, I. F. Laurenson, A. Barrett, F. Drobniewski, N. D. McCarthy, L. F. Anderson, I. Abubakar, H. L. Thomas, P. Monk, E. G. Smith, A. S. Walker, D. W. Crook, T. E. Peto and C. P. Conlon (2014). "Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study." The lancet. Respiratory medicine **2**(4): 285-292.

Wallace, R. J., Jr., B. A. Brown and D. E. Griffith (1998). "Nosocomial outbreaks/pseudo-outbreaks caused by nontuberculous mycobacteria." Annual Review of Microbiology **52**: 453-490.

Wallace, R. J., Jr., A. Meier, B. A. Brown, Y. Zhang, P. Sander, G. O. Onyi and E. C. Bottger (1996). "Genetic basis for clarithromycin resistance among isolates of

Mycobacterium chelonae and Mycobacterium abscessus." Antimicrobial Agents and Chemotherapy **40**(7): 1676-1681.

Walters, S. B., E. Dubnau, I. Kolesnikova, F. Laval, M. Daffe and I. Smith (2006). "The Mycobacterium tuberculosis PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis." Molecular Microbiology **60**(2): 312-330.

Wanner, R. M., D. Castor, C. Guthlein, E. C. Bottger, B. Springer and J. Jiricny (2009). "The uracil DNA glycosylase UdgB of Mycobacterium smegmatis protects the organism from the mutagenic effects of cytosine and adenine deamination." Journal of Bacteriology **191**(20): 6312-6319.

Warren, R. M., T. C. Victor, E. M. Streicher, M. Richardson, N. Beyers, N. C. Gey van Pittius and P. D. van Helden (2004). "Patients with active tuberculosis often have different strains in the same sputum specimen." American Journal of Respiratory and Critical Care Medicine **169**(5): 610-614.

Watson, S. J., M. R. Welkers, D. P. Depledge, E. Coulter, J. M. Breuer, M. D. de Jong and P. Kellam (2013). "Viral population analysis and minority-variant detection using short read next-generation sequencing." Philosophical transactions of the Royal Society of London. Series B, Biological sciences **368**(1614): 20120205.

Weiss, C. H. and J. Glassroth (2012). "Pulmonary disease caused by nontuberculous mycobacteria." Expert review of respiratory medicine **6**(6): 597-612; quiz 613.

Wendt, S. L., K. L. George, B. C. Parker, H. Gruft and J. O. Falkinham, 3rd (1980). "Epidemiology of infection by nontuberculous Mycobacteria. III. Isolation of potentially pathogenic mycobacteria from aerosols." The American review of respiratory disease **122**(2): 259-263.

Werngren, J. and S. E. Hoffner (2003). "Drug-susceptible Mycobacterium tuberculosis Beijing genotype does not develop mutation-conferred resistance to rifampin at an elevated rate." Journal of Clinical Microbiology **41**(4): 1520-1524.

Wertman, R., M. Miller, P. Groben, D. S. Morrell and D. A. Culton (2011). "Mycobacterium bolletii/Mycobacterium massiliense furunculosis associated with pedicure footbaths: a report of 3 cases." Archives of Dermatology **147**(4): 454-458.

WHO (2012). Global Tuberculosis Report 2012, World Health Organisation.

WHO (2013). Global Tuberculosis Report 2013, World Health Organisation.

Wiehlmann, L., G. Wagner, N. Cramer, B. Siebert, P. Gudowius, G. Morales, T. Kohler, C. van Delden, C. Weinel, P. Slickers and B. Tummler (2007). "Population structure of Pseudomonas aeruginosa." Proceedings of the National Academy of Sciences of the United States of America **104**(19): 8101-8106.

Workentine, M. L., C. D. Sibley, B. Glezerson, S. Purighalla, J. C. Norgaard-Gron, M. D. Parkins, H. R. Rabin and M. G. Surette (2013). "Phenotypic heterogeneity of Pseudomonas aeruginosa populations in a cystic fibrosis patient." PloS one **8**(4): e60225.

World Health Organization (2012). Leprosy. http://www.who.int/mediacentre/factsheets/fs101/en/. **101**.

World Health Organization (2013). Buruli ulcer (Mycobacterium ulcerans infection). http://www.who.int/mediacentre/factsheets/fs199/en/index.html. **199**.

Yamachika, S., C. Sugihara, Y. Kamai and M. Yamashita "Correlation between penicillin-binding protein 2 mutations and carbapenem resistance in Escherichia coli." J Med Microbiol **62**(Pt 3): 429-436.

Ypma, R. J., T. Donker, W. M. van Ballegooijen and J. Wallinga (2013). "Finding evidence for local transmission of contagious disease in molecular epidemiological datasets." PloS one **8**(7): e69875.

Zagordi, O., R. Klein, M. Daumer and N. Beerenwinkel (2010). "Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies." Nucleic Acids Research **38**(21): 7400-7409.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome research **18**(5): 821-829.

Zhang, H., D. Li, L. Zhao, J. Fleming, N. Lin, T. Wang, Z. Liu, C. Li, N. Galwey, J. Deng, Y. Zhou, Y. Zhu, Y. Gao, S. Wang, Y. Huang, M. Wang, Q. Zhong, L. Zhou, T. Chen, J. Zhou, R. Yang, G. Zhu, H. Hang, J. Zhang, F. Li, K. Wan, J. Wang, X. E. Zhang and L. Bi (2013). "Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance." Nature Genetics **45**(10): 1255-1260.

Zhang, Y., M. A. Yakrus, E. A. Graviss, N. Williams-Bouyer, C. Turenne, A. Kabani and R. J. Wallace, Jr. (2004). "Pulsed-field gel electrophoresis study of Mycobacterium abscessus isolates previously affected by DNA degradation." Journal of Clinical Microbiology **42**(12): 5582-5587.

Zumla, A., P. Mwaba, J. Huggett, N. Kapata, D. Chanda and J. Grange (2009). "Reflections on the white plague." The Lancet infectious diseases **9**(3): 197-202.

9. Appendix

9.1. Chapter 1

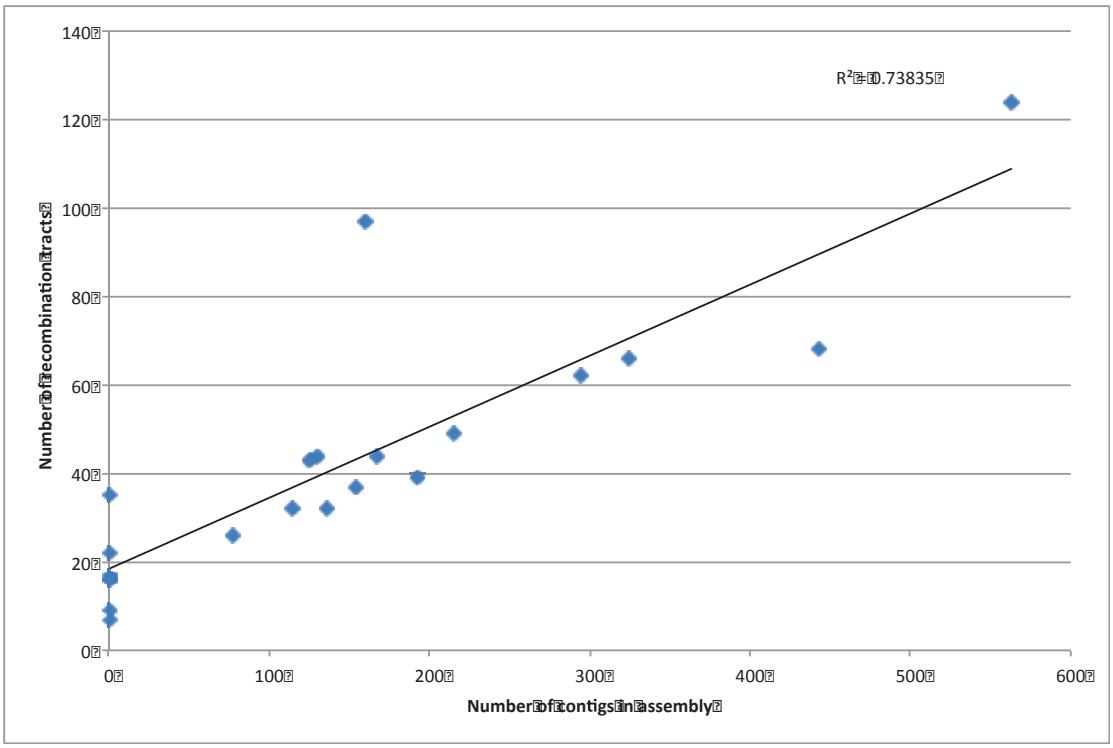


Figure 47 - Correlation between assembly quality and number of recombination events reported in Namouchi *et. al.* 2012. The number of contigs in each assembly was found to correlate well with the number of recombination events detected, with very few being detected in complete genomes (1 contig). This suggests that the recombination events detected are likely to be artifactual, due to errors found at the end of contigs. This finding is part of a manuscript in preparation.

9.2. Chapter 2

Appendix table 1 – Metadata associated with analysis in Chapter 2. ERS accession number refers to data on the European Nucleotide Archive

Strain name	ERS number	Month of collection	Year of collection
2007-874	ERS016362	5	2007
2007-875	ERS016363	5	2007
A00301419	ERS007637	8	2003
A00400237	ERS007638	2	2004
A00400885	ERS007642	4	2004
A00401093	ERS007644	6	2004
A00401371	ERS013436	8	2004
A00500149	ERS007646	1	2005
A09900565	ERS007636	3	1999
A09901284	ERS007640	6	1999
B00001265	ERS007651	7	2000
B00600458	ERS007652	2	2006

B00801080	ERS007655	6	2008
B09601851	ERS007647	9	1996
B09601926	ERS007649	10	1996
B09701572	ERS007654	7	1997
C00401055	ERS007659	6	2004
C09400866	ERS007656	3	1994
C09601061	ERS007657	5	1996
D00000071	ERS007663	12	1999
D00000538	ERS007664	3	2000
D00501624	ERS007661	10	2005
D00700688	ERS007665	4	2007
E09700455	ERS007667	2	1997
E09701106	ERS007670	6	1997
E09701191	ERS007668	5	1997
E09701891	ERS007672	8	1997
E09801032	ERS007673	3	1998
EE0502463	ERS016332	4	2005
EE0503811	ERS016342	?	2005
EE0506178	ERS016360	9	2005
EE0506784	ERS016343	11	2005
F00001728	ERS007675	8	2000
F00002017	ERS013466	11	2000
F00201178	ERS007677	5	2002
F09900854	ERS007674	3	1999
G00000978	ERS007683	5	2000
G00001233	ERS007684	6	2000
G00101715	ERS007686	6	2001
G09801536	ERS007679	6	1998
G09901357	ERS007681	7	1999
H00200930	ERS007690	4	2002
H09601792	ERS007688	9	1996
I00200020	ERS007694	12	2001
I00300092	ERS013491	12	2002
I00300818	ERS013492	4	2003
I00601066	ERS013493	6	2006
I09400880	ERS007692	3	1994
I09400970	ERS007695	4	1994
I09601853	ERS013489	9	1996
I09702330	ERS007693	11	1997
I09800059	ERS013490	12	1997
J09400698	ERS013418	1	1994
J09401337	ERS013419	5	1994
J09402203	ERS013420	10	1994
J09500385	ERS013421	1	1995
J09500392	ERS013422	1	1995
J09602061	ERS013423	11	1996
J09700140	ERS013424	12	1996
J09700837	ERS013425	4	1997
J09701231	ERS013426	5	1997
J09701362	ERS016341	6	1997
J09701593	ERS013427	8	1997
J09701920	ERS013428	8	1997
J09800554	ERS013429	2	1998

J09800628	ERS013430	3	1998
J09800745	ERS013431	3	1998
J09900137	ERS013432	12	1998
J09900221	ERS013433	12	1998
J09900583	ERS013434	2	1999
J09902073	ERS013435	9	1999
K00101539	ERS007701	6	2001
K00200539	ERS007702	3	2002
K00201415	ERS007708	8	2002
K00500041	ERS007703	12	2004
K00500133	ERS007710	1	2005
K00700186	ERS016350	1	2007
K09600911	ERS007697	4	1996
K09700503	ERS007704	3	1997
K09900646	ERS007699	2	1999
K09901478	ERS013495	7	1999
K09901972	ERS007706	9	1999
L00201493	ERS007711	8	2002
L00401786	ERS007712	11	2004
M00018020	ERS013502	10	1993
M00102039	ERS007724	10	2001
M00102058	ERS007714	10	2001
M00201208	ERS007726	6	2002
M09400328	ERS007713	12	1993
M09400374	ERS013504	1	1994
M09400397	ERS007716	1	1994
M09400937	ERS007718	3	1994
M09401471	ERS007720	6	1994
M09402043	ERS013503	9	1994
M09500081	ERS013505	12	1994
M09502139	ERS007722	11	1995
N00000064	ERS016355	11	1999
N00000221	ERS007728	1	2000
N00000626	ERS007730	4	2000
N00001761	ERS007731	9	2000
N00001996	ERS013514	9	2000
N00200616	ERS013515	4	2002
N00302086	ERS013517	10	2003
N00400624	ERS016352	4	2004
N00400960	ERS013501	5	2004
N00401131	ERS013511	6	2004
N00500166	ERS013507	1	2005
N00500759	ERS013512	5	2005
N00501578	ERS016357	9	2005
N00501741	ERS016358	11	2005
N00701874	ERS013513	11	2007
N00800620	ERS016356	4	2008
N00800922	ERS013510	5	2008
N09501026	ERS013516	3	1995
N09900612	ERS016353	3	1999
N09900805	ERS013506	4	1999
N09900855	ERS013508	4	1999
N09901297	ERS007734	6	1999

N09901607	ERS016354	7	1999
N09901618	ERS007733	8	1999
N09901774	ERS007737	8	1999
N09901829	ERS013509	8	1999
N09902308	ERS007739	11	1999
O00100977	ERS013440	5	2001
O00200127	ERS016333	12	2001
O00401640	ERS013445	10	2004
O09401777	ERS013437	8	1994
O09501089	ERS013446	4	1995
O09501127	ERS013447	5	1995
O09501596	ERS013441	7	1995
O09700898	ERS013438	3	1997
O09700920	ERS013439	4	1997
O09701563	ERS013442	7	1997
O09801157	ERS013443	5	1998
O09801359	ERS013444	6	1998
P00301553	ERS013449	7	2003
P00401210	ERS016361	7	2004
P00401720	ERS013457	10	2004
P00601339	ERS007743	6	2006
P00601463	ERS007741	6	2006
P09501076	ERS013450	4	1995
P09501164	ERS013451	4	1995
P09501258	ERS016334	5	1994
P09501329	ERS013448	6	1995
P09501583	ERS013452	5	1994
P09501896	ERS013453	9	1995
P09502221	ERS013454	11	1995
P09600222	ERS013455	10	1995
P09701236	ERS013456	6	1997
Q00501124	ERS016337	6	2005
Q09600228	ERS013458	12	1995
Q09600750	ERS013459	4	1996
Q09701888	ERS016336	9	1997
Q09802318	ERS016335	10	1998
Q09900060	ERS013460	11	1998
Q09900131	ERS013461	12	1998
Q09900809	ERS013462	4	1999
Q09901139	ERS013463	5	1999
Q09901557	ERS013464	7	1999
Q09901859	ERS013465	9	1999
R00018412	ERS013474	11	1993
R00301759	ERS013468	10	2003
R00401166	ERS013473	7	2004
R09401256	ERS013469	5	1994
R09401839	ERS013470	7	1994
R09401899	ERS016338	9	1994
R09402296	ERS016339	11	1994
R09500613	ERS013471	2	1995
R09500614	ERS013472	2	1995
R09601315	ERS013467	6	1996
R09801836	ERS016340	8	1998

S00600143	ERS016344	1	2006
S00600351	ERS013475	1	2006
SAWC-507	ERS016364	?	?
T00200958	ERS013480	5	2002
T09601545	ERS013476	8	1996
T09800777	ERS013481	3	1998
T09800928	ERS013482	3	1998
T09900050	ERS013477	11	1998
T09900518	ERS013478	2	1999
T09901127	ERS013479	6	1999
U09801187	ERS016345	5	1998
U09801472	ERS013483	6	1998
V09501561	ERS016346	7	1995
V09601310	ERS013484	6	1996
W09900252	ERS016348	1	1999
W09900339	ERS016347	1	1999
W09900422	ERS013485	1	1999
W09900423	ERS013486	1	1999
W09900434	ERS013487	1	1999
W09900442	ERS013488	2	1999
X00800532	ERS013494	3	2008
X00800994	ERS016349	5	2008
Y00015573	ERS013496	7	1992
Y00015898	ERS013497	9	1992
Y09400034	ERS013498	12	1993
Y09500720	ERS013499	9	1992
Z09600572	ERS016351	3	1996
Z09901262	ERS013500	6	1999

9.3. Chapter 3

Appendix table 2 - Metadata associated with ReMoxTB study.

Patient number	Accession of primary isolate (a)	Accession of secondary isolate (b)	Time between episodes (weeks)	SNP distance	Site
2	ERS075344	ERS075345	36	64	Cape Town, SA
3	ERS075347	ERS075346	36	3	Stellenbosch, SA
4	ERS075348	ERS075349	48	6	Stellenbosch, SA
5	ERS075350	ERS075351	26	0	Stellenbosch, SA
6	ERS075352	ERS075353	26	0	Cape Town, SA
7	ERS075354	ERS075355	36	0	Stellenbosch, SA
8	ERS075356	ERS075357	26	48	Stellenbosch, SA
9	ERS075358	ERS075359	26	0	Stellenbosch, SA
10	ERS075360	ERS075361	48	1419	Cape Town, SA
11	ERS075363	ERS075362	26	0	Stellenbosch, SA

12	ERS075365	ERS075364	36	0	Stellenbosch, SA
13	ERS075366	ERS075367	36	1	Stellenbosch, SA
14	ERS075369	ERS075368	60	1340	Cape Town, SA
15	ERS075371	ERS075370	26	1364	Stellenbosch, SA
16	ERS075372	ERS075373	26	0	Cape Town, SA
17	ERS075374	ERS075375	26	0	Cape Town, SA
18	ERS075376	ERS075377	48	0	Stellenbosch, SA
19	ERS075378	ERS075379	36	0	Stellenbosch, SA
20	ERS075381	ERS075380	48	0	Stellenbosch, SA
21	ERS075382	ERS075383	36	1	Stellenbosch, SA
22	ERS075384	ERS075385	26	0	Cape Town, SA
23	ERS075386	ERS075387	36	0	Stellenbosch, SA
24	ERS075388	ERS075389	36	0	Cape Town, SA
25	ERS075390	ERS075391	48	0	Cape Town, SA
26	ERS075392	ERS075393	36	0	Cape Town, SA
27	ERS075394	ERS075395	36	0	Stellenbosch, SA
28	ERS075396	ERS075397	36	0	Stellenbosch, SA
29	ERS075398	ERS075399	36	0	Cape Town, SA
30	ERS075400	ERS075401	26	0	Stellenbosch, SA
31	ERS075402	ERS075403	26	0	Stellenbosch, SA
32	ERS124347	ERS124348	26	0	Durban, SA
33	ERS124349	ERS124350	60	0	Durban, SA
34	ERS124351	ERS124352	36	0	Durban, SA
35	ERS124353	ERS124354	17	1306	Durban, SA
36	ERS124355	ERS124356	28	898	Durban, SA
37	ERS124357	ERS124358	48	1207	Durban, SA
38	ERS124359	ERS124360	60	767	Durban, SA
40	ERS124363	ERS124364	36	0	Durban, SA
41	ERS124365	ERS124366	28	2	Johannesburg, SA
42	ERS124367	ERS124368	37	1	Brits, SA
43	ERS124369	ERS124370	36	1	Brits, SA
44	ERS124371	ERS124372	48	0	Brits, SA
45	ERS124374	ERS124373	17	1	Kuala Lumpur, MY
47	ERS124378	ERS124377	17	0	Kuala Lumpur, MY
48	ERS124379	ERS124380	36	0	Nonthaburi, TH
49	ERS124381	ERS124382	36	2	Nonthaburi, TH
50	ERS124383	ERS124384	36	0	Nonthaburi, TH

9.4. Chapter 4

Appendix table 3 - Number of assembled PE, PE-PGRS and PPE genes

Gene	Number of isolates for which is was assembled (total=48)	Gene	Number of isolates for which is was assembled (total=48)	Gene	Number of isolates for which is was assembled (total=48)
PE1	48	PE_PGRS1	48	PPE1	48
PE2	48	PE_PGRS2	48	PPE2	48
PE3	48	PE_PGRS3	7	PPE3	48
PE4	48	PE_PGRS4	40	PPE4	48
PE5	48	PE_PGRS5	48	PPE5	48
PE6	48	PE_PGRS6	47	PPE7	33
PE7	48	PE_PGRS7	48	PPE9	48
PE8	48	PE_PGRS8	48	PPE10	48
PE9	48	PE_PGRS9	44	PPE11	48
PE10	48	PE_PGRS10	44	PPE12	48
PE11	48	PE_PGRS11	48	PPE13	45
PE12	48	PE_PGRS13	48	PPE14	48
PE13	48	PE_PGRS14	48	PPE15	48
PE14	48	PE_PGRS15	48	PPE16	26
PE15	48	PE_PGRS16	48	PPE17	48
PE16	48	PE_PGRS17	7	PPE18	13
PE17	48	PE_PGRS18	19	PPE19	16
PE18	46	PE_PGRS19	36	PPE20	48
PE19	48	PE_PGRS20	17	PPE21	48
PE20	48	PE_PGRS21	46	PPE22	48
PE22	48	PE_PGRS22	38	PPE23	48
PE23	48	PE_PGRS23	48	PPE24	33
PE24	48	PE_PGRS24	48	PPE25	9
PE25	48	PE_PGRS25	48	PPE26	46
PE26	48	PE_PGRS26	48	PPE27	0
PE27A	39	PE_PGRS27	10	PPE28	48
PE27	48	PE_PGRS28	10	PPE29	48
PE29	9	PE_PGRS29	48	PPE30	48
PE31	48	PE_PGRS30	48	PPE31	48
PE32	48	PE_PGRS31	48	PPE32	48
PE33	48	PE_PGRS32	48	PPE33	48
PE34	48	PE_PGRS33	48	PPE34	20

PE35	48	PE_PGRS34	48	PPE35	48
PE36	48	PE_PGRS35	48	PPE36	48
		PE_PGRS36	48	PPE37	48
		PE_PGRS37	48	PPE38	6
		PE_PGRS38	48	PPE39	48
		PE_PGRS39	48	PPE40	14
		PE_PGRS40	48	PPE41	48
		PE_PGRS41	48	PPE42	48
		PE_PGRS42	48	PPE43	48
		PE_PGRS43	48	PPE44	48
		PE_PGRS44	48	PPE45	48
		PE_PGRS45	13	PPE46	9
		PE_PGRS46	48	PPE47	20
		PE_PGRS47	48	PPE49	44
		PE_PGRS48	48	PPE50	26
		PE_PGRS50	33	PPE51	48
		PE_PGRS51	48	PPE52	48
		PE_PGRS52	45	PPE53	20
		PE_PGRS53	45	PPE54	5
		PE_PGRS54	18	PPE55	43
		PE_PGRS55	11	PPE56	43
		PE_PGRS58	47	PPE57	21
		PE_PGRS57	0	PPE58	34
		PE_PGRS59	48	PPE59	20
		PE_PGRS60	48	PPE60	39
		PE_PGRS61	48	PPE61	48
		PE_PGRS62	48	PPE62	48
		PE_PGRS63	48	PPE63	48
		wag22	19	PPE64	48
				PPE65	48
				PPE66	46
				PPE68	48
				PPE69	46

9.5. Chapter 5

Appendix table 4 - Metadata associated with *M. abscessus* studies. Run accessions are specified as there were sometimes multiple runs for the same sample. Subspecies designations were inferred from whole genome data.

Strain name	Run accession	Sample accession	Month	Year	Subspecies
1_b	ERR119107	ERS075661	6	2008	abscessus sensu stricto
1_c	ERR115000	ERS075527	7	2008	abscessus sensu stricto
1_d	ERR115004	ERS075531	9	2008	abscessus sensu stricto
1_e	ERR115005	ERS075532	9	2008	abscessus sensu stricto
1_f	ERR115040	ERS075567	3	2010	abscessus sensu stricto
1_g	ERR115046	ERS075573	4	2010	abscessus sensu stricto
1_h	ERR115079	ERS075606	7	2010	abscessus sensu stricto
10_a	ERR115063	ERS075590	11	2010	bollettii
11_a	ERR114976	ERS075503	1	2008	abscessus sensu stricto
11_b	ERR114996	ERS075523	1	2008	abscessus sensu stricto
11_c	ERR114972	ERS075499	3	2009	abscessus sensu stricto
11_d	ERR115043	ERS075570	7	2009	abscessus sensu stricto
11_e	ERR115026	ERS075553	7	2009	abscessus sensu stricto
12_a	ERR115108	ERS075635	7	2008	massiliense
12_b	ERR115003	ERS075530	7	2008	massiliense
12_c	ERR115019	ERS075546	3	2009	massiliense
12_d	ERR115023	ERS075550	4	2009	massiliense
12_e	ERR119106	ERS075660	8	2009	massiliense
12_f	ERR115036	ERS075563	2	2010	massiliense
12_g	ERR115050	ERS075577	5	2010	massiliense
13_a	ERR115007	ERS075534	10	2008	massiliense
13_b	ERR115034	ERS075561	1	2010	massiliense
14_a	ERR115012	ERS075539	12	2008	massiliense
14_b	ERR115013	ERS075540	12	2008	massiliense
14_c	ERR119104	ERS075658	12	2008	massiliense
14_d	ERR119105	ERS075659	1	2009	massiliense
14_e	ERR115021	ERS075548	4	2009	massiliense
14_f	ERR115071	ERS075598	1	2011	massiliense
14_g	ERR115087	ERS075614	2	2011	massiliense
14_h	ERR115088	ERS075615	2	2011	massiliense
14_i	ERR115075	ERS075602	3	2011	massiliense
15_a	ERR115070	ERS075597	12	2010	abscessus sensu stricto
16_a	ERR115020	ERS075547	3	2009	mixed
16_b	ERR115025	ERS075552	5	2009	mixed
16_c	ERR119102	ERS075656	12	2009	mixed
16_d	ERR119103	ERS075657	3	2010	mixed
17_a	ERR115022	ERS075549	4	2009	massiliense
18_a	ERR115030	ERS075557	8	2009	massiliense
18_b	ERR119095	ERS075649	8	2009	massiliense
19_a	ERR119096	ERS075650	9	2009	massiliense
19_b	ERR115101	ERS075628	10	2009	massiliense
19_c	ERR119084	ERS075638	11	2009	massiliense
19_d	ERR119082	ERS075636	11	2009	massiliense
19_e	ERR119088	ERS075642	11	2009	massiliense
19_f	ERR119085	ERS075639	1	2010	massiliense
19_g	ERR119083	ERS075637	1	2010	massiliense
19_h	ERR119089	ERS075643	1	2010	massiliense

19_i	ERR115107	ERS075634	3	2010	massiliense
19_j	ERR115044	ERS075571	3	2010	massiliense
19_k	ERR119086	ERS075640	4	2010	massiliense
19_l	ERR119087	ERS075641	5	2010	massiliense
19_m	ERR115051	ERS075578	5	2010	massiliense
19_n	ERR115055	ERS075582	7	2010	massiliense
19_o	ERR115080	ERS075607	8	2010	massiliense
19_p	ERR115059	ERS075586	8	2010	massiliense
19_q	ERR115083	ERS075610	9	2010	massiliense
19_r	ERR115069	ERS075596	12	2010	massiliense
2_a	ERR114970	ERS075497	11	2007	massiliense
2_aa	ERR115053	ERS075580	6	2010	massiliense
2_ab	ERR115057	ERS075584	8	2010	massiliense
2_ac	ERR119108	ERS075662	11	2010	massiliense
2_ad	ERR115086	ERS075613	2	2011	massiliense
2_b	ERR114971	ERS075498	11	2007	massiliense
2_c	ERR114973	ERS075500	12	2007	massiliense
2_d	ERR115098	ERS075625	12	2007	massiliense
2_e	ERR115093	ERS075620	1	2008	massiliense
2_f	ERR114980	ERS075507	2	2008	massiliense
2_g	ERR114981	ERS075508	2	2008	massiliense
2_h	ERR114982	ERS075509	2	2008	massiliense
2_i	ERR114995	ERS075522	3	2008	massiliense
2_j	ERR114988	ERS075515	3	2008	massiliense
2_k	ERR114989	ERS075516	3	2008	massiliense
2_l	ERR114992	ERS075519	3	2008	massiliense
2_m	ERR114990	ERS075517	3	2008	massiliense
2_n	ERR114991	ERS075518	3	2008	massiliense
2_o	ERR114998	ERS075525	6	2008	massiliense
2_p	ERR115001	ERS075528	7	2008	massiliense
2_q	ERR115009	ERS075536	10	2008	massiliense
2_r	ERR115014	ERS075541	12	2008	massiliense
2_s	ERR115015	ERS075542	1	2009	massiliense
2_t	ERR115016	ERS075543	1	2009	massiliense
2_u	ERR119090	ERS075644	6	2009	massiliense
2_v	ERR119092	ERS075646	8	2009	massiliense
2_w	ERR119091	ERS075645	9	2009	massiliense
2_x	ERR119093	ERS075647	10	2009	massiliense
2_y	ERR119094	ERS075648	12	2009	massiliense
2_z	ERR115099	ERS075626	3	2010	massiliense
20_a	ERR115105	ERS075632	11	2009	massiliense
20_b	ERR115037	ERS075564	2	2010	massiliense
20_c	ERR115045	ERS075572	3	2010	massiliense
20_d	ERR115048	ERS075575	5	2010	massiliense
20_e	ERR115078	ERS075605	7	2010	massiliense
20_f	ERR115056	ERS075583	7	2010	massiliense
20_g	ERR115060	ERS075587	9	2010	massiliense
20_h	ERR115068	ERS075595	12	2010	massiliense
21_a	ERR115031	ERS075558	11	2009	abscessus sensu stricto
21_b	ERR115032	ERS075559	12	2009	abscessus sensu stricto
21_c	ERR115033	ERS075560	12	2009	abscessus sensu stricto
21_d	ERR115047	ERS075574	5	2010	abscessus sensu stricto
22_a	ERR119101	ERS075655	3	2010	massiliense

22_b	ERR115038	ERS075565	2	2010	massiliense
22_c	ERR115042	ERS075569	3	2010	massiliense
22_d	ERR115058	ERS075585	8	2010	massiliense
22_e	ERR115082	ERS075609	9	2010	massiliense
22_f	ERR115064	ERS075591	11	2010	massiliense
22_g	ERR115067	ERS075594	12	2010	massiliense
22_h	ERR115073	ERS075600	2	2011	massiliense
22_i	ERR115076	ERS075603	3	2011	massiliense
23_a	ERR115039	ERS075566	2	2010	abscessus sensu stricto
24_a	ERR115049	ERS075576	5	2010	abscessus sensu stricto
25_a	ERR115052	ERS075579	6	2010	massiliense
26_a	ERR115054	ERS075581	7	2010	abscessus sensu stricto
27_a	ERR115081	ERS075608	8	2010	massiliense
28_a	ERR115061	ERS075588	10	2010	massiliense
28_b	ERR115062	ERS075589	11	2010	massiliense
28_c	ERR115065	ERS075592	11	2010	massiliense
28_d	ERR115084	ERS075611	11	2010	massiliense
28_e	ERR115085	ERS075612	1	2011	massiliense
28_f	ERR115074	ERS075601	2	2011	massiliense
29_a	ERR115072	ERS075599	2	2011	massiliense
29_b	ERR115089	ERS075616	3	2011	massiliense
29_c	ERR115092	ERS075619	4	2011	massiliense
3_a	ERR114977	ERS075504	1	2008	bollettii
3_b	ERR114978	ERS075505	1	2008	bollettii
3_c	ERR115106	ERS075633	1	2008	bollettii
3_d	ERR114979	ERS075506	1	2008	bollettii
3_e	ERR114997	ERS075524	5	2008	bollettii
3_f	ERR115002	ERS075529	7	2008	bollettii
3_g	ERR115008	ERS075535	10	2008	bollettii
3_h	ERR115017	ERS075544	2	2009	bollettii
3_i	ERR115095	ERS075622	2	2009	bollettii
3_j	ERR115027	ERS075554	7	2009	bollettii
3_k	ERR115028	ERS075555	7	2009	bollettii
3_l	ERR115029	ERS075556	7	2009	bollettii
3_m	ERR115103	ERS075630	12	2009	bollettii
3_n	ERR119097	ERS075651	5	2010	bollettii
30_a	ERR115090	ERS075617	3	2011	massiliense
31_a	ERR115091	ERS075618	3	2011	abscessus sensu stricto
31_b	ERR115077	ERS075604	4	2011	abscessus sensu stricto
4_a	ERR114967	ERS075494	11	2007	abscessus sensu stricto
4_b	ERR115035	ERS075562	11	2007	abscessus sensu stricto
5_a	ERR115094	ERS075621	12	2007	abscessus sensu stricto
5_b	ERR114975	ERS075502	1	2008	abscessus sensu stricto
5_c	ERR115096	ERS075623	1	2008	abscessus sensu stricto
5_d	ERR114983	ERS075510	2	2008	abscessus sensu stricto
5_e	ERR114993	ERS075520	4	2008	abscessus sensu stricto
5_f	ERR114994	ERS075521	4	2008	abscessus sensu stricto
5_g	ERR115006	ERS075533	9	2008	abscessus sensu stricto
5_h	ERR115010	ERS075537	11	2008	abscessus sensu stricto
5_i	ERR115011	ERS075538	11	2008	abscessus sensu stricto
5_j	ERR115018	ERS075545	2	2009	abscessus sensu stricto
5_k	ERR115024	ERS075551	5	2009	abscessus sensu stricto
5_l	ERR119100	ERS075654	10	2009	abscessus sensu stricto

5_m	ERR119098	ERS075652	12	2009	abscessus sensu stricto
5_n	ERR115097	ERS075624	1	2010	abscessus sensu stricto
5_o	ERR119099	ERS075653	2	2010	abscessus sensu stricto
5_p	ERR115066	ERS075593	12	2010	abscessus sensu stricto
6_a	ERR114968	ERS075495	11	2007	abscessus sensu stricto
7_a	ERR115041	ERS075568	10	2007	abscessus sensu stricto
7_b	ERR114965	ERS075492	10	2007	abscessus sensu stricto
7_c	ERR114966	ERS075493	10	2007	abscessus sensu stricto
7_d	ERR114984	ERS075511	10	2007	abscessus sensu stricto
8_a	ERR114987	ERS075514	12	2007	massiliense
8_b	ERR115100	ERS075627	12	2007	massiliense
9_a	ERR114986	ERS075513	11	2007	abscessus sensu stricto
9_b	ERR115102	ERS075629	11	2007	abscessus sensu stricto
9_c	ERR114974	ERS075501	12	2007	abscessus sensu stricto
9_d	ERR115104	ERS075631	12	2007	abscessus sensu stricto

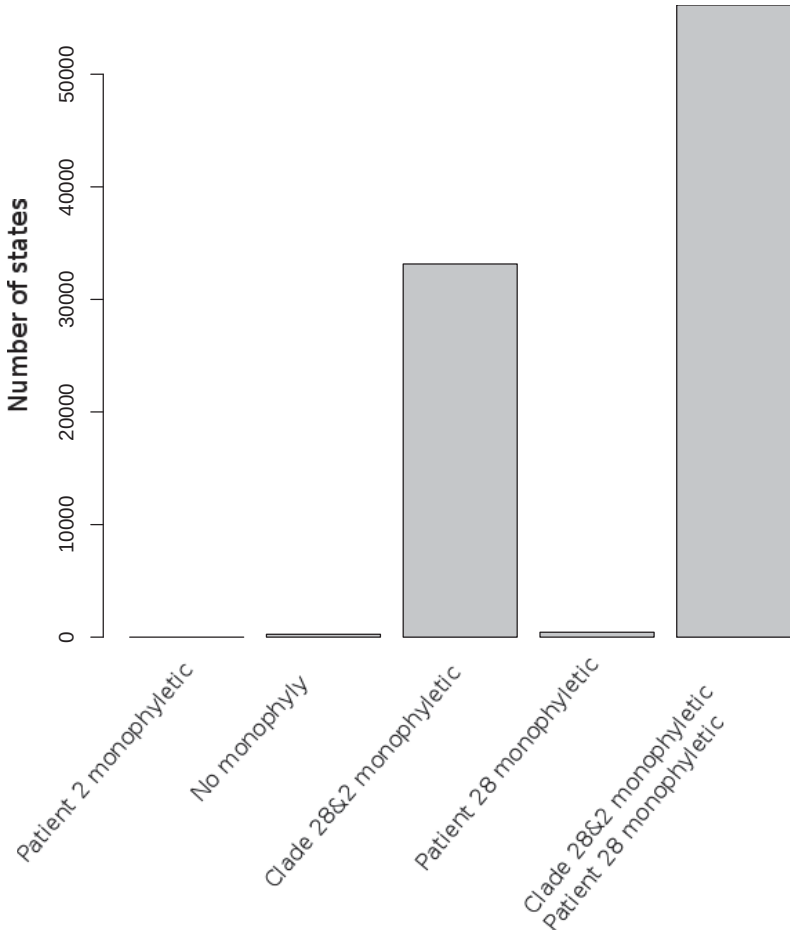


Figure 48 - Additional BEAST run of monophyly test reported in section 5.3.5

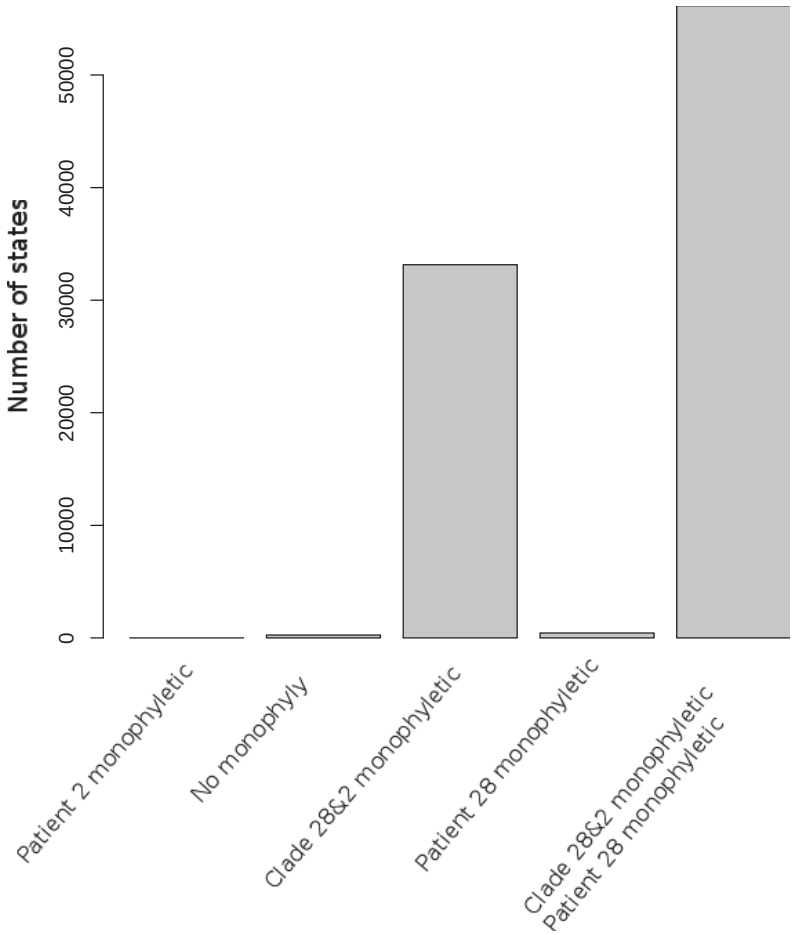


Figure 49. Additional BEAST run of monophyly test reported in section 5.3.5

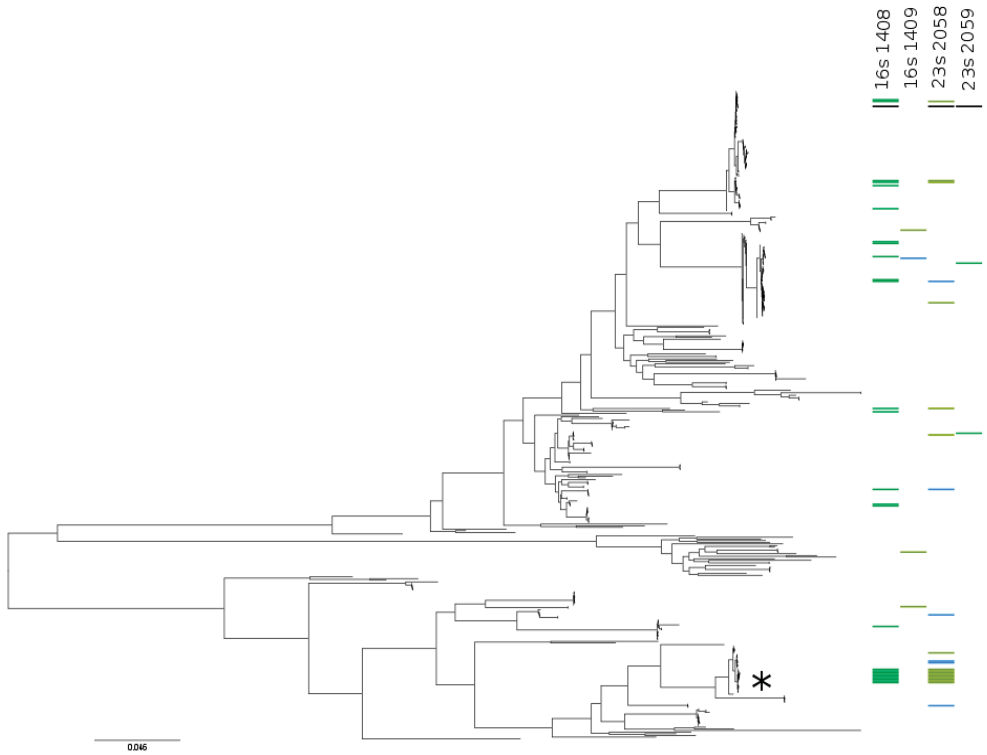


Figure 50. Amikacin and clarithromycin resistance in preliminary analysis of global collection of *M. abscessus* isolates. ML tree of first isolates collected from ~300 patients as part of global *M. abscessus* diversity project (unpublished data) constructed using RAxML (Stamatakis 2006). The alleles of loci known to confer resistance to amikacin or clarithromycin were extracted from the mapping data, and presented next to the corresponding isolate on the tree. Different colours represent different nucleotides. This demonstrates that resistance is rarely sustained in the population, except in the case of the Papworth *M. a. massiliense* transmission clusters (indicated with star).

9.6. Chapter 6

Appendix table 5 - Samples used in within-patient diversity sample and their colony morphotype. R= rough, S= Smooth

Sample	Colony morphotype	Sample	Colony morphotype
11a	R	22h	S
11b	R	22i	S
11c	R	28a	R
11d	R	28b	R

11e	R	28c	R
12a	S	28d	R
12b	S	28e	R
12c	S	28f	R
12d	S	2a	R
12e	S	2aa	R
12f	S	2ac	R
12g	S	2ad	R
13a	R	2b	R
13b	R	2c	R
14b	S	2d	R
14c	S	2e	R
14d	S	2f	R
14e	S	2g	R
14g	R	2h	R
14h	S	2i	R
14i	R	2j	R
18a	R	2k	R
18b	R	2l	R
19a	S	2m	R
19b	S	2n	R
19c	S	2o	R
19d	S	2p	S
19e	S	2r	R
19f	S	2s	R
19g	S	2t	R
19h	S	2u	R
19i	S	2v	S
19j	S	2w	R
19k	S	2x	R
19l	S	2y	R
19m	S	2z	R
19n	S/R	3a	R
19o	S	3b	R
19p	S	3c	R
19q	S	3d	R
19r	S	3e	R
1a	S	3f	R
1c	R	3g	R
1d	S	3h	R
1e	S	3i	R
1f	R	3j	R
1g	R	3l	R
20a	S/R	3m	R
20b	S/R	3n	R
20c	S	5a	S

20d	S	5b	S
20e	S	5d	S
20f	S	5e	S
20g	S/R	5f	S
20h	S/R	5g	S
22a	S	5h	S
22b	S	5j	S/R
22c	S	5k	S
22e	R	5m	R
22f	S	5n	S
22g	S	5o	S
		5p	S

9.6.1. List of possible antibiotic resistance associated genes in *M. abscessus*

MAB_0035c
MAB_0036c
MAB_0313c
MAB_0327
MAB_0408c
MAB_0519
MAB_0951
MAB_1257
MAB_2000
MAB_2179
MAB_2297
MAB_2359
MAB_2875
MAB_3165c
MAB_3167c
MAB_3637c
MAB_3681
MAB_4395
MAB_4482
MAB_4901c
MAB_4910c
MAB_0330
MAB_0414
MAB_0696c
MAB_1114
MAB_1312
MAB_1386
MAB_1387
MAB_1870
MAB_2179

MAB_2314c
MAB_2833
MAB_2875
MAB_4006
MAB_4231
MAB_4755c
MAB_4800
MAB_4805
MAB_4947
MAB_r5051
MAB_r5052

Appendix table 6 - Frequency of recurrent minority variants identified in patient 2 and 28

Sample	Date	MAB_0140c	MAB_0173	MAB_0179	upstream of MAB_0408c	MAB_0416c	MAB_0477	MAB_0674	MAB_1057	intergenic	MAB_1683	MAB_2618	MAB_2742	MAB_3072	upstream of MAB_3360c & MAB_3361
2a	11/30/07	0.00	0.51	0.06	0.47	0.50	0.00	0.00	0.14	0.03	0.50	0.55	0.00	0.00	0.00
2b	11/30/07	0.47	0.97	0.10	0.00	0.81	0.00	0.00	0.64	0.06	0.97	0.81	0.00	0.64	0.00
2c	12/13/07	0.00	0.95	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.99	0.94	0.00	0.00	0.00
2d	12/13/07	0.24	0.99	0.00	0.00	0.96	0.00	0.64	0.97	0.00	0.98	0.96	0.00	0.23	0.00
2e	1/28/08	0.21	0.82	0.00	0.06	0.59	0.11	0.00	0.30	0.00	0.92	0.72	0.00	0.16	0.00
2f	2/8/08	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00
2g	2/8/08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2h	2/9/08	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
2i	3/5/08	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
2j	3/15/08	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2k	3/18/08	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00
2l	3/18/08	0.00	0.99	0.98	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
2m	3/19/08	0.00	0.96	0.00	0.00	0.92	0.40	0.00	0.39	0.00	0.96	0.93	0.00	0.00	0.00
2n	3/19/08	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.99	0.00	1.00	1.00	0.00	0.00	0.00
2o	6/4/08	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.99	0.00	1.00	0.99	0.00	0.00	0.00
2p	7/29/08	0.00	0.97	0.00	0.00	0.96	0.00	0.99	0.99	0.00	0.97	0.95	0.00	0.00	0.00
2r	12/17/08	0.00	0.79	0.00	0.00	0.81	0.79	0.00	0.00	0.00	0.89	0.87	0.00	0.00	0.10
2s	1/7/09	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.99	0.00	1.00	1.00	0.00	0.00	0.00
2t	1/21/09	0.00	1.00	0.00	0.00	1.00	0.99	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
2u	6/2/09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
2v	8/25/09	0.32	0.99	0.39	0.00	0.36	0.00	0.06	0.34	0.00	1.00	0.31	0.00	0.31	0.00
2w	9/30/09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99
2x	10/17/09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

2y	12/8/09	0.00	0.12	0.00	0.00	0.16	0.15	0.00	0.00	0.00	0.08	0.16	0.07	0.00	0.33
2z	3/31/10	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96
2aa	6/11/10	0.00	0.12	0.12	0.36	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.35	0.00	0.00
2ac	11/12/10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2ad	2/6/11	0.00	0.08	0.00	0.00	0.05	0.06	0.00	0.00	0.00	0.07	0.03	0.00	0.00	0.83
28a	26/10/2010	0.00	0.99	0.00	0.00	0.99	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
28b	08/11/2010	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
28c	09/11/2010	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
28d	24/11/2010	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
28e	23/01/2011	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.99	1.00	0.00	0.00	0.00
28f	28/02/2011	0.00	1.00	0.00	0.00	0.99	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00