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).

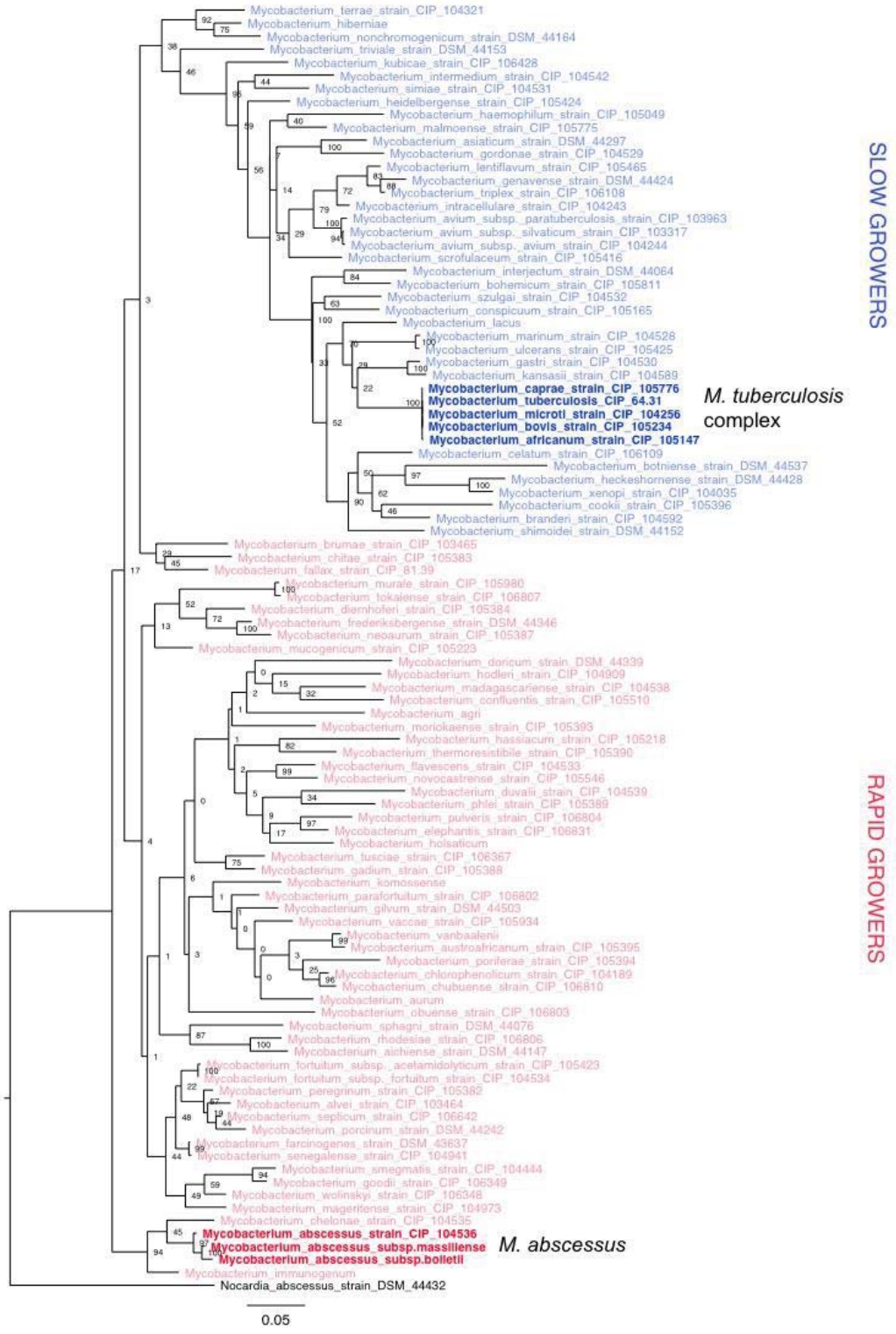

Figure 1 - Phylogeny of mycobacteria based on concatenated *Sod*, *Hsp65* and 16S sequences. Sequence accessions obtained from Devulder *et al.* (Devulder, Perouse de Montclos *et al.* 2005). Aligned with Muscle (Edgar 2004) and tree built with RAxML (Stamatakis 2006). Bootstrap support obtained from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

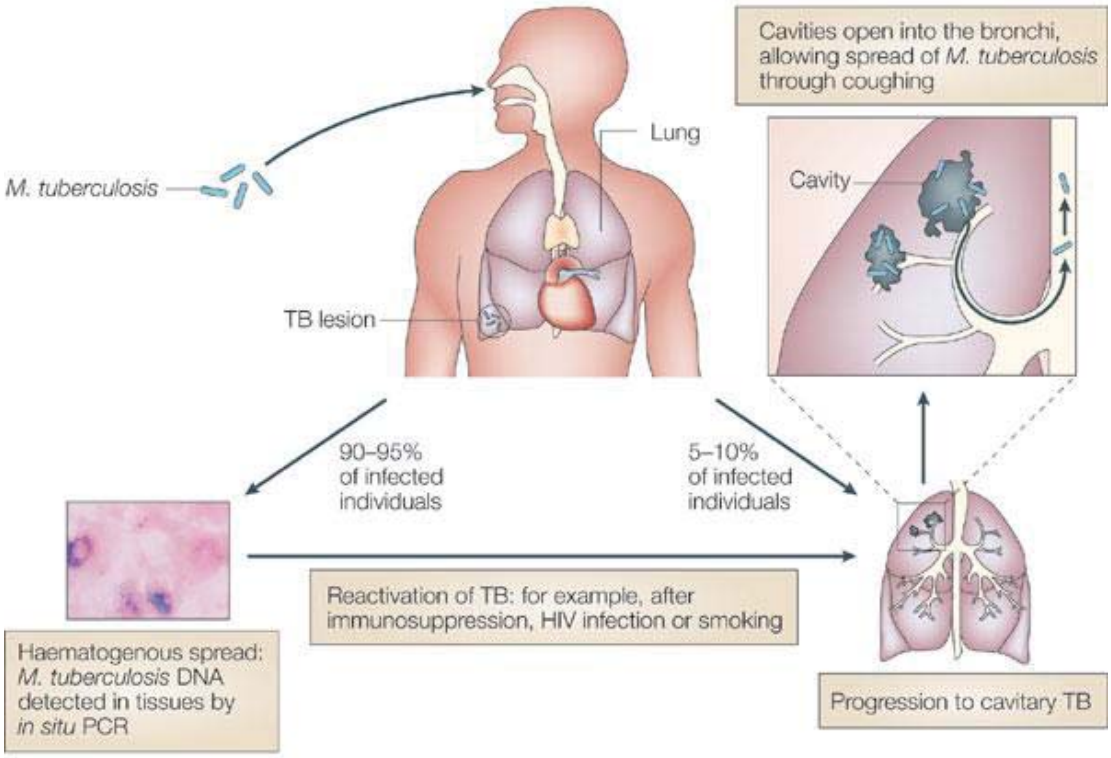

Figure 2 - Phases of human tuberculosis. After inhalation of the bacteria, there is a blood-borne stage where the immune system attempts to control the infection. In 5-10% of individuals this will lead to active or cavitary tuberculosis, which can shed *M. tuberculosis* allowing ongoing transmission 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).

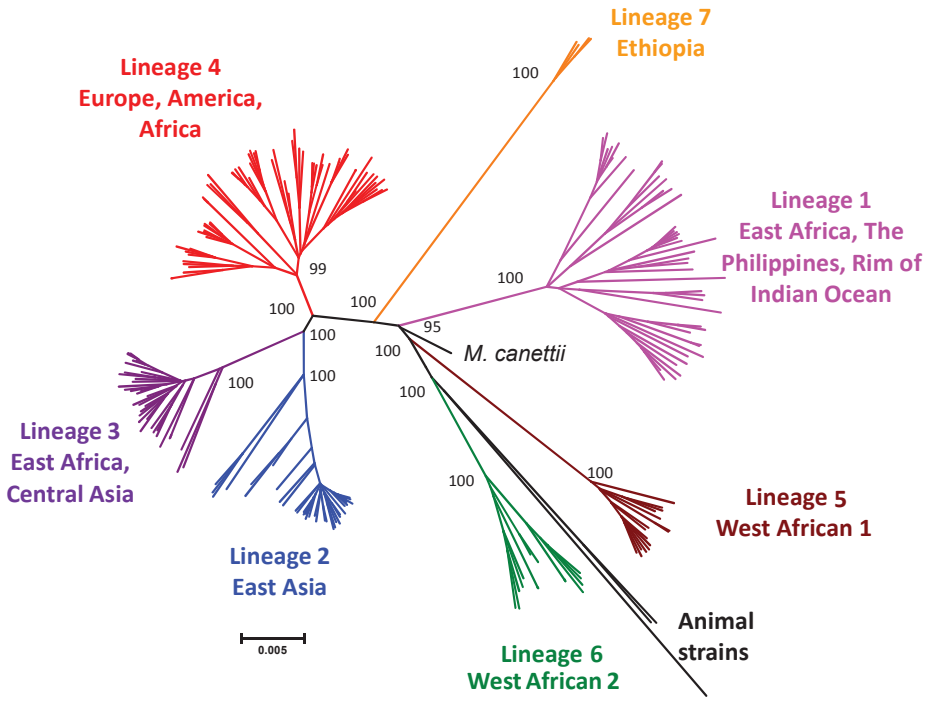

Figure 3 – Maximum likelihood phylogeny of the M. tuberculosis complex (MTBC) based on whole genome sequences. Adapted by permission from Macmillan Publishers Ltd: Nature Genetics, Comas *et. al.*, copyright 2013. Bootstrap support for the lineages are shown. The branch leading to *M. canettii* has been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

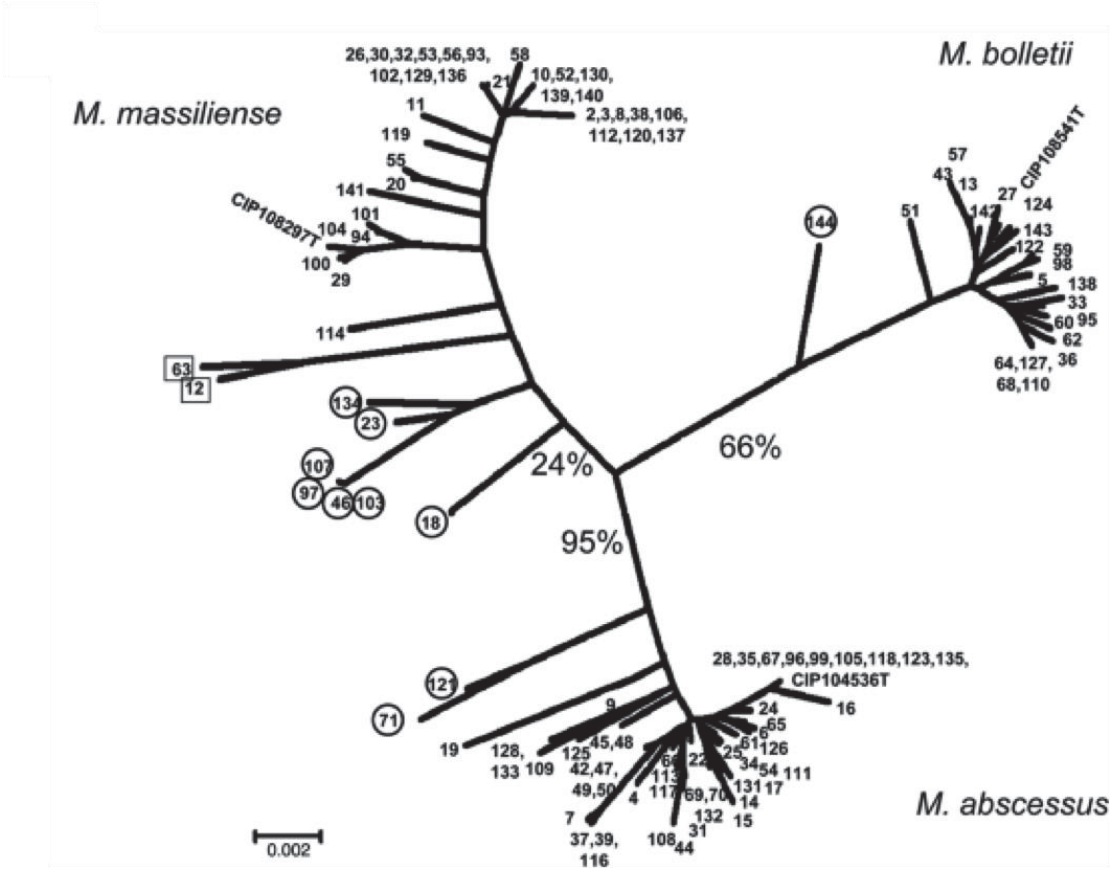

Figure 4 - MLSA based phylogeny of *M. abscessus* strains. Percentages indicate bootstrap support. Figure adapted from Macheras *et al*. (2011) with permission from the American Society for Microbiology.

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.