# 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.