Evolutionary Genomics of Cystic Fibrosis and Nosocomial Pathogens of the *Mycobacterium abscessus* species complex

Isobel Everall

St Catharine's College

University of Cambridge

Wellcome Sanger Institute

September 2018

This dissertation is submitted for the degree of Doctor of Philosophy

Evolutionary Genomics of Cystic Fibrosis and Nosocomial Pathogens of the *Mycobacterium abscessus* species complex

Isobel Everall

Summary:

The *Mycobacterium abscessus* species complex (MABSC) consists of three subspecies, *M. a. abscessus*, *M. a. massiliense* and *M. a. bolletii*. All three of these subspecies are capable of causing opportunistic pulmonary and skin and soft tissue infections in immunocompromised individuals. Infections caused by the MABSC are particularly serious in people with underlying lung conditions such as Cystic Fibrosis (CF) as these organisms are highly antibiotic resistant and the currently available treatment is toxic. This has resulted in treatment of MABSC infections failing in up to 50% of cases. Given the increasing prevalence globally of infections caused by the MABSC, particularly in individuals with CF, there is increased urgency to improve the treatment available for MABSC infections. The overall aim of this project was to use genomic analyses to increase our understanding of how the MABSC has evolved to become an opportunistic pathogen, which in turn could potentially uncover promising targets for the development of novel antibiotics.

Genomic analysis of the MABSC has already shown that lineages of the MABSC are capable of indirect person to person transmission and the extent to which transmission contributed to the increasing prevalence of MABSC in people with CF became evident when the MABSC global population structure was determined. This showed that 70% of the MABSC isolates from people with CF were attributed to lineages made up of densely clustered isolates, where acquisition *via* indirect person-to-person transmission, as opposed to from the environment, was more likely.

This thesis used population genomic approaches to i) investigate the genetic factors that drove the emergence of the most prevalent MABSC lineages, ii) look for evidence of convergence after the clonal expansion of these lineages to understand how the MABSC was continuing to adapt and spread amongst people with CF and iii) to examine the within host evolution of these pathogens to uncover how these environmental organisms were adapting to the CF lung. This thesis also used whole genome sequencing to explore the

largest known outbreak of MABSC infections, a post-surgical wound infection epidemic in Brazil.

Through this research the emergence of the most prevalent MABSC lineages were found to be driven by increased opportunity, probably due to the increased number of people with CF surviving longer, as opposed to the acquisition of a common genetic determinant. Not enough signal was detected after the clonal expansion of the most prevalent MABSC lineages to come to strong conclusions about genetic factors driving their continuing expansion, but strong evidence of convergent evolution was detected between MABSC isolates evolving over time within the host. The MABSC was shown to be potentially using a similar central regulatory network in response to environmental cues from the phagosome to that of *M. tuberculosis*. The investigation into the epidemic of post-surgical wound infections in Brazil showed that a single *M. a. massiliense* lineage was introduced into Brazil just prior to the initial outbreak and that this lineage subsequently spread, through several waves of transmission, to multiple cities in geographically distant areas of Brazil. This highlighted how the MABSC was capable of long distance transmission and emphasized the potential of the MABSC as a nosocomial pathogen capable of causing large scale outbreaks.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically stated in the text.

The work presented in this thesis is not substantially the same as any that I have submitted, or, is being submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution.

This thesis does not exceed the prescribed word limit stipulated by the Biological Sciences Degree committee

Abstract

The *Mycobacterium abscessus* species complex (MABSC) consists of three subspecies, *M. a. abscessus*, *M. a. massiliense* and *M. a. bolletii*. All three of these subspecies are capable of causing opportunistic pulmonary and skin and soft tissue infections in immunocompromised individuals. Infections caused by the MABSC are particularly serious in people with underlying lung conditions such as Cystic Fibrosis (CF) as these organisms are highly antibiotic resistant and the currently available treatment is toxic. This has resulted in treatment of MABSC infections failing in up to 50% of cases. Given the increasing prevalence globally of infections caused by the MABSC, particularly in individuals with CF, there is increased urgency to improve the treatment available for MABSC infections. The overall aim of this project was to use genomic analyses to increase our understanding of how the MABSC has evolved to become an opportunistic pathogen, which in turn could potentially uncover promising targets for the development of novel antibiotics.

Genomic analysis of the MABSC has already shown that lineages of the MABSC are capable of indirect person to person transmission and the extent to which transmission contributed to the increasing prevalence of MABSC in people with CF became evident when the MABSC global population structure was determined. This showed that 70% of the MABSC isolates from people with CF were attributed to lineages made up of densely clustered isolates, where acquisition *via* indirect person-to-person transmission, as opposed to from the environment, was more likely.

This thesis used population genomic approaches to i) investigate the genetic factors that drove the emergence of the most prevalent MABSC lineages, ii) look for evidence of convergence after the clonal expansion of these lineages to understand how the MABSC was continuing to adapt and spread amongst people with CF and iii) to examine the within host evolution of these pathogens to uncover how these environmental organisms were adapting to the CF lung. This thesis also used whole genome sequencing to explore the largest known outbreak of MABSC infections, a post-surgical wound infection epidemic in Brazil.

Through this research the emergence of the most prevalent MABSC lineages were found to be driven by increased opportunity, probably due to the increased number of people with CF surviving longer, as opposed to the acquisition of a common genetic determinant. Not enough signal was detected after the clonal expansion of the most prevalent MABSC

lineages to come to strong conclusions about genetic factors driving their continuing expansion, but strong evidence of convergent evolution was detected between MABSC isolates evolving over time within the host. The MABSC was shown to be potentially using a similar central regulatory network in response to environmental cues from the phagosome to that of *M. tuberculosis*. The investigation into the epidemic of post-surgical wound infections in Brazil showed that a single *M. a. massiliense* lineage was introduced into Brazil just prior to the initial outbreak and that this lineage subsequently spread, through several waves of transmission, to multiple cities in geographically distant areas of Brazil. This highlighted how the MABSC was capable of long distance transmission and emphasized the potential of the MABSC as a nosocomial pathogen capable of causing large scale outbreaks.

Acknowledgements

Having dreamt of being at the stage of writing my acknowledgements for months I'm suddenly speechless. This project would not have been possible without Julian Parkhill, who gave me this opportunity and has been on hand to guide and advise me through it. Not only does he seem to have no boundaries to his knowledge, he has also been unfailingly kind and understanding through all the trials and tribulations as I have tried to expand mine. I'm also very much indebted to Andres Floto. My project always seemed to progress or take a new turn after a meeting with him. Together, Julian and Andres make a great team and I feel very fortunate to have worked with them both. I'm also grateful to Nick Thompson who together with Julian and Andres formed my thesis committee and provided useful advice and guidance. I'm also thankful to the CF Trust for funding my PhD.

I'm very grateful to Simon Harris (and his script directory) who has always been on hand to offer invaluable advice and guidance. My fellow abscessus survivors, Josie and Dani, have both been incredibly helpful and supportive and really kept me going throughout the project. I'm also extremely grateful to everyone in the Parkhill and Floto labs for all their help and morale boosting over the past four years. On the Parkhill side; James, Sophia, Leo, Becca (who's always been kind enough to listen to my rants), Steph, Gerry, Sam, Chris, Neil, Crispin, Narender, Nicole and John (who put up with sitting next to me for 3 of the 4 years) all provided invaluable help and laughs, whilst ChaP, Sophie, Mini, Lucas, Karen, Dani and Josie always made me feel welcome in the Floto lab.

Finally, thanks to my friends and family for their unwavering support through this challenge. To PhD14 – I can't believe we're at the end.. we've done it!. To Catz MCR – thanks (but not from my liver) for all the entertainment, it has massively contributed making my time in Cambridge unforgettable. The top floor crew – thanks for the great memories and friendships, Erin – thanks for all the G&Ts and being a great friend (Fynn - thanks for making many of them!). To Dami – thanks for listening to my practice presentations, and always believing I would get to the end. Rhys, a fellow PhD sufferer, thanks for all the morale boosting messages. Caitlin, thanks for the support and some crazy times and yes I've hopefully finally finished. Finally, to my mum, dad and sister, who I know are extremely glad this is over, there are no words, I hope I can repay you in the future.

Contents

List of abbreviations

AAI	Amino acid Identity
ANI	Average nucleotide identity
ASL	Airway surface liquid
BCAA	Branched chain amino acids
BP	Biological Process
CC	Cellular Component
CF	Cystic Fibrosis
CFTR	Cystic Fibrosis Transmembrane Conductance Regulator
COG	Cluster of Orthologous Groups
CRP	Cyclic-AMP receptor binding protein
DCC	Dominant circulating clone
DDH	DNA-DNA hybridization
ddNTPs	Dideoxynucleotide triphosphates
dNTPs	Deoxynucleotide triphosphates
GO-terms	Gene ontology terms
GPLs	Glycopeptidolipids
GTA	Glutaraldehyde
IPDR	Interpulse duration ratio
KEGG	Kyto Encyclopedia of Genes and Genomes
LAM	Lipoarabinomannan
LCA	Last common ancestor
MABSC	*Mycobacterium abscessus* species complex
MAC	*Mycobacterium avium* complex
M. a. abscessus	*Mycobacterium abscessus* subspecies *abscessus*
M. a. bolletii	*Mycobacterium abscessus* subspecies *bolletii*
M. a. massiliense	*Mycobacterium abscessus* subspecies *massiliense*
mce	Mammalian cell entry
MF	Molecular function
MFS	Major facilitator superfamily protein
MODQV	Modification quality value
msp	Mycobacterium smegmatis porin
MST	Minimum Spanning Tree

NTM	Nontuberculous Mycobacteria
RGM	Rapidly growing mycobacteria
RM	Restriction modification
RPKM	Reads per kilobase million
SCFM	Synthetic Cystic Fibrosis medium
SGM	Slow growing mycobacteria
SCID	Severe combined immunodeficiency
SMRT	Single molecule real time sequencing
SSTI	Skin and soft tissue infection
TCS	Two component system
T7SS	Type VII secretion system
VNTR	Variable number tandem repeat
WGS	Whole Genome Sequencing

1. Introduction

1. Introduction

1.1 Cystic Fibrosis

1.1.1 A brief history of Cystic Fibrosis

"'Woe to that child who tastes salty when kissed on the forehead. He is bewitched and soon must die" (as quoted in (1)). An Irish proverb dating back to the late fifteenth century hints that there was an awareness of Cystic Fibrosis (CF) long before the first descriptions of the condition were published in the 1930s. 'Fibrocystic disease of the pancreas' and 'cystic fibromatosis with bronchiectasis' were the terms first used to distinguish the combined observations of abnormal pancreatic function, malnutrition, poor growth and chronic lung infections from digestive conditions such as coeliac disease and to suggest a single disease entity was responsible for the co-occurrence of these symptoms (2, 3).

The start of the antibiotic era in the 1940s was the first medical intervention that enabled a significant number of people with CF to survive beyond infancy as they began to be treated with first generation antibiotics penicillin, terramycin and aeruomycin (4, 5). In the same decade a heatwave in New York led to the discovery of abnormally high salt concentrations in the sweat of people with CF, which both suggested that ion imbalances were associated with CF and was the foundation of the now commonly used diagnosis technique the 'sweat test' (6). The 1940s was also the decade in which CF was first proposed to be a genetic disorder that was inherited in an autosomal recessive manner, although it wouldn't be until the late 1980s that the gene responsible, the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR), was identified on chromosome 7 at position q31.2 (7-10). By this time, epithelial dysfunction and specifically chloride ions being unable to cross the epithelial cell membrane had been identified as the basic defect in CF (11-13). This was confirmed in the 1990s when CFTR was found to function as a chloride ion channel, bringing together the previous six decades of research (11-13).

The functional 1480 amino acid CFTR gene encodes a chloride ion channel, expressed on the surface of epithelial cells, consisting of two membrane spanning domains, made up of six transmembrane helices, two nuclear binding domains, nuclear binding domain 1 (NBD-1) and nuclear binding domain 2 (NBD-2), and a regulatory region (9, 14, 15). To date nearly 2000 CFTR mutations affecting the synthesis (Del1078T), trafficking (DelF508), gating (G551D), conductance (R117H) and the amount of functional protein that reaches the epithelial membrane (A455E) have been described, although not all have been proven to cause CF (16, 17). The distribution of epithelial cells across many parts of the body means

the loss or dysfunction of CFTR has a wide range of implications, with symptoms affecting the pancreas, liver, sweat glands, sinuses, intestines, male fertility tract and lungs.

There are approximately 100,000 CF sufferers worldwide, the majority of whom are of Caucasian ethnicity, where CF is the most common genetically inherited disorder (18, 19). The reason for the increased CF allele prevalence in Caucasians has yet to be conclusively proven, with the most convincing hypotheses suggesting that a heterozygous CFTR phenotype provided a selective advantage against infectious diseases. These have been hypothesized to be either typhoid fever, because *Salmonella typhi* has been shown to bind to CFTR, or Cholera, as being heterozygous for CFTR has been suggested to reduce gastrointestinal fluid loss (20, 21). For similar reasons, CFTR heterozygosity has also been proposed to have been potentially beneficial to people with lactose intolerance associated diarrhea (22). However, Poolman and Galvani, argue that *Mycobacterium tuberculosis* is the infectious agent with the strongest correlation between clinical and molecular evidence of CFTR heterozygosity providing a selective advantage against the disease and a historical event, the 17th century Tuberculosis pandemic, that could have provided the selective pressure to increase the CF allele frequency in the Caucasian population (23). Lubinsky (2012) expanded this hypothesis further suggesting that a combination of the incidence of tuberculosis and hypertension in the population as well as vitamin D deficiency, temperature and altitude explain the variation in the CF allele frequency globally (24).

Currently in the UK there are approximately 10,460 people living with CF, with a person born with CF approximately 1 in 2,500 births (25). A combination of improved nutrition, the availability of antibiotics, novel and more efficient delivery mechanisms for these antibiotics, intensive physiotherapy and an effective mucolytic to aid mucus clearance in the lungs as well as stringent cross-infection prevention protocols has resulted in a dramatically increased life expectancy for people with CF (26) (Figure 1).

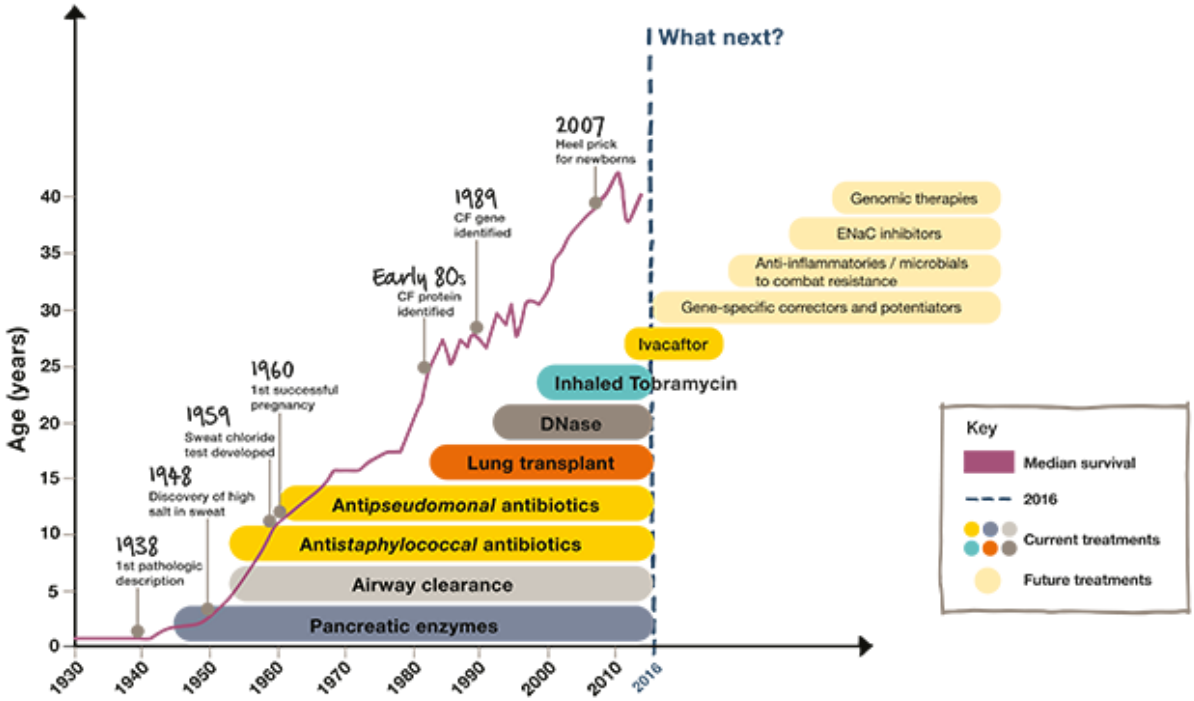


Figure 1: Increasing median survival age of people with Cystic Fibrosis

Since the first pathological description of CF in 1938 the median survival age for people with CF has increased from not surviving beyond infancy to 47 (25). The introduction of several treatments has contributed to this increased median survival age, however as more people with CF survive longer they are exposed to an increasing array of opportunistic pathogens. Figure reproduced with permission from the Cystic Fibrosis Trust.

However, despite great hope for gene therapy and the gradual introduction of drugs targeting specific CFTR mutations (27-29), there is still no cure for CF and it is still a life limiting condition, with a median survival age of 47 in the UK (25).The biggest contributor to this limited life expectancy, estimated to be responsible for 80% to 95% of morbidity and mortality associated with CF, is the loss of lung function caused by chronic lung infections (19, 30, 31).

1.1.2 Lung Pathophysiology in Cystic Fibrosis

In the lungs, the loss of the CFTR from the membrane of the airway epithelial cells results in an imbalance of chloride (Cl^-) and sodium (Na^+) ions passing across the airway epithelial membrane (Figure 2) (26). This results in i) an osmotic pressure that draws water into the epithelial cells causing the dehydration of the airway surface liquid (ASL), ii) the lowering of the pH of the ASL and iii) the destabilization of mucins. Together these features give rise to the thick and sticky sputum characteristic of CF lungs (Figure 2) (26). The nature of this sputum causes one of the key innate defense mechanisms of the lungs, mucociliary

clearance, to become ineffective as the cilia are unable to beat and therefore cannot clear invading inhaled microorganisms (26). The inability to clear microorganisms is compounded by the ASL in people with CF being rich in DNA, amino acids and iron which provide a source of nutrients to the colonizing microorganisms (32).

As well as the loss of the CFTR ion channel causing significant changes to the ASL which are beneficial to invading microorganisms, the loss of the CFTR channel itself also impairs the host immune response (32). The CFTR ion channel has also been shown to be capable of binding pathogenic microorganisms such as *Pseudomonas aeruginosa*, causing them to be engulfed into the epithelial cell and lysed (32). It is also the channel by which antioxidants, such as glutathione (GTH) and thiocyanate (SCN^-), are introduced into the airway, and consequently the host is unable to fully control the oxidative stress response in the airways, resulting in more significant lung damage (32). The levels of nitric oxide synthase and nitric oxide are also reduced by the loss of CFTR which consequently leads to neutrophil killing being less effective (32). Glycolipids on the surface of the airway epithelial cells also become acylated due to the loss/dysfunction of CFTR, increasing bacterial adherence to the airway epithelial surface, and thus increasing the chance of an infection taking hold (31).The effect of the loss or dysfunction of CFTR on the lung microenvironment, as described above, shows why people with CF are so susceptible to acquiring lung infections.

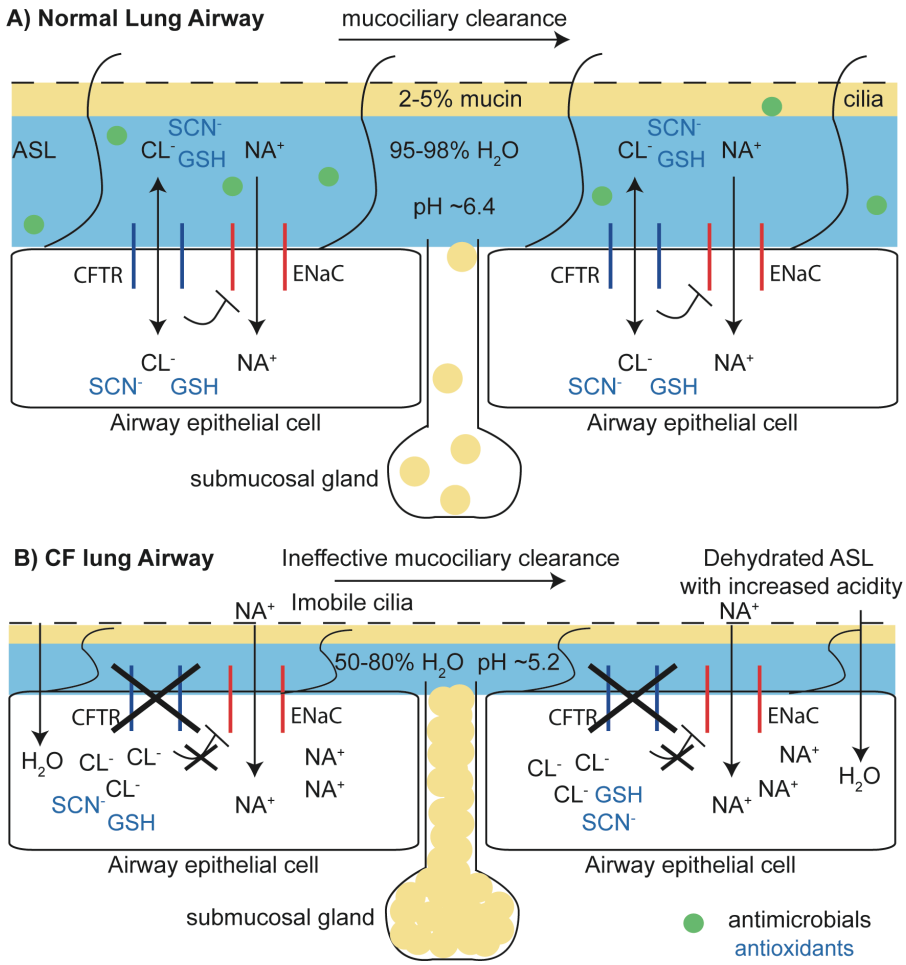


Figure 2: Pathophysiology of the Cystic Fibrosis Lung

In the normal lung airway (A), with a functioning CFTR, the balance of ions crossing the epithelial membrane is such that the airway surface liquid (ASL) is at the right level of hydration and acidity for optimum mucociliary clearance. The antioxidants Thiocyanate (SCN^-), and glutathione (GSH) are also secreted through CFTR in the healthy lung. The loss of CFTR (B), causes an imbalance of ions across the epithelial membrane, with CFTR no longer negatively regulating the ENaC channel, leading to the creation of an osmotic pressure drawing water into the cells and dehydrating the ASL. The ASL in CF lungs is therefore thick and sticky which causes mucocillary clearance to be ineffective. The innate immune response is also affected with the pH of CF ASL reduced and antioxidants secreted through CFTR no longer secreted.

Studies have shown that, whilst a person with CF is born with the same lung anatomy as that of an unaffected individual, on average within the first year of life they acquire a lung infection (32). A cycle of infection then begins, which causes chronic lung inflammation and eventually leads to bronchiectasis and a fatal decline in lung function (33). The causal agents of these infections include bacteria, fungal and viral pathogens, however, whilst the CF lung is a polymicrobial environment, a few species commonly cause the majority of infections in

people with CF (34). Although, as people with CF live longer and novel treatments are developed the epidemiology of CF lung infections is changing (35, 36).

1.1.3 Traditional Cystic Fibrosis Pathogens

The most common CF pathogens globally are *Pseudomonas aeruginosa*, *Staphylococcus aureus* and *Haemophilus influenzae* (Figure 3A) (19). The epidemiology of the CF lung infections changes with age. Commonly, the first infectious agents isolated from the patients with CF are *Haemophilus influenzae* and *Staphylococcus aureus,* but as people with CF get older, *Pseudomonas aeruginosa* becomes the most prevalent cause of infection (Figure 3B) (19). Other pathogens that cause a significant disease burden in CF include the organisms of the *Burkholderia cepacia* complex, the fungal pathogen *Aspergillus fumigatus*, *Ralstonia* species, and *Pandoraea* species (37, 38).

As well as the CF lung being a hospitable environment for these pathogens, the pathogens themselves have adapted to thrive in this environment, making the infections harder to treat. CF pathogens have been shown to adapt to the CF lung by acquiring resistance to antibiotics, switching from synthesizing amino acids to using the abundance of amino acids in CF sputum, incurring pathoadaptive mutations in genes that enhance their ability to form biofilms and thus protect themselves from both antibiotics and the host's immune system and by acquiring a hypermutator phenotype (39-42). Many of these adaptations, whilst making the pathogens less virulent, contribute to the difficulty in eradicating the infection and consequently the slow decline in lung function and eventual death.

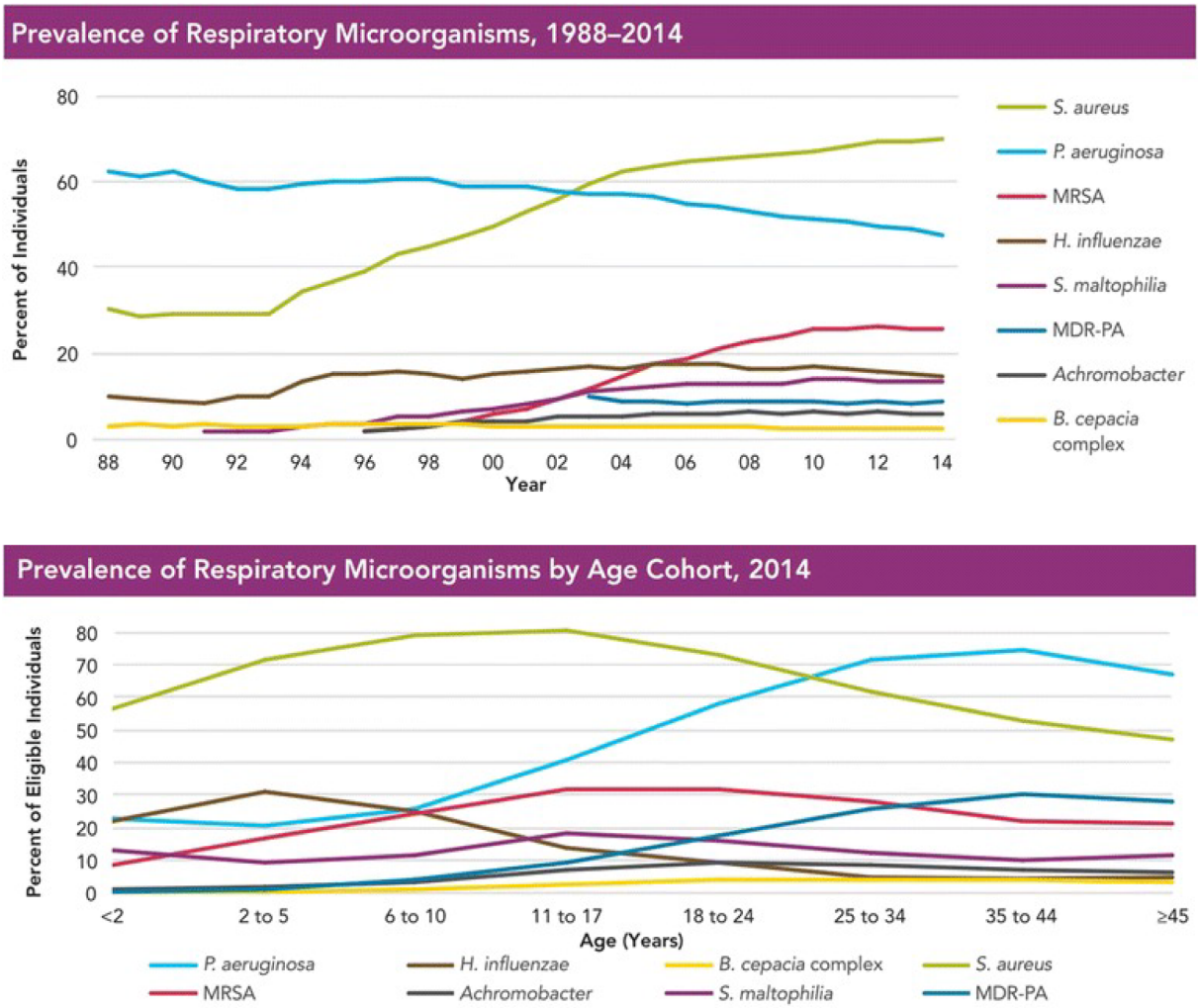


Figure 3: Prevalence of traditional CF pathogens

The top panel shows the changes in prevalence of CF pathogens between 1988 and 2014 and the bottom panels shows the prevalence of CF pathogens by age. These graphs show the emergence of novel pathogens infecting people with CF, including *Stenotrophomonas maltophilia,* Methicillin Resistant *Staphylococcus aureus* (MRSA) and *Achromoboacter* spp. Both these graphs are based upon data collected by the United States CF patient registry, with the figure originally published by Bhagirath et al. 2016 (19).

Initially, the vast majority of people with CF were believed to acquire pathogens from the environment, with evidence of transmission generally only observed between sibling pairs (43, 44). However, an outbreak at a CF holiday camp caused by the *Burkholderia cepacia* complex proved conclusively that transmission of pathogens between people with CF did occur (45). Subsequently, further transmissible lineages of CF pathogens have emerged, with epidemic lineages of *Pseudomonas aeruginosa* (*e.g.* Liverpool epidemic strain, Australian epidemic strains, Manchester epidemic strain) and further transmissible *Burkholderia cepacia* complex lineages (e.g. ET12) observed (46-50). Transmissible

lineages of CF pathogens are commonly responsible for the highest proportion of infections with that species and in some cases have been shown to have increased virulence and be associated with rapid clinical decline and increased mortality (48, 51). The introduction of strict infection control protocols has reduced transmission between people with CF, although person to person transmission has not been eradicated. Furthermore, as segregation of CF patients in hospitals is usually only applied if a particular pathogen has been proven to be transmissible, unexpected direct or indirect person to person transmission involving other species can occur.

Despite the adaptation of CF pathogens to the CF lung, the median life expectancy of people with CF has improved (see above). However, the increasing number of people with CF living longer combined with new treatments opening up novel niches within the CF lung has resulted in more opportunistic pathogens having the opportunity to come into contact with the CF lung and consequently further opportunistic pathogens are emerging as a serious threat to people with CF (52).

1.1.4 Emerging Cystic Fibrosis Pathogens

Over the past decade the epidemiology of the CF lung has changed. Whilst the traditional CF pathogens are still responsible for a large proportion of lung infections in people with CF, other pathogens, some well-known and some rare, are emerging that pose new treatment challenges. These include methicillin resistant *Staphylococcus aureus* (MRSA), which has increased in prevalence most evidently in US CF centers, pathogens more commonly associated with nosocomial infections such as *Stenotrophomonas maltophilia* and the relatively rare opportunistic pathogens *Achromobacter xylosoxidans* and Non-tuberculous mycobacteria (NTM) (Figure 4) (38).

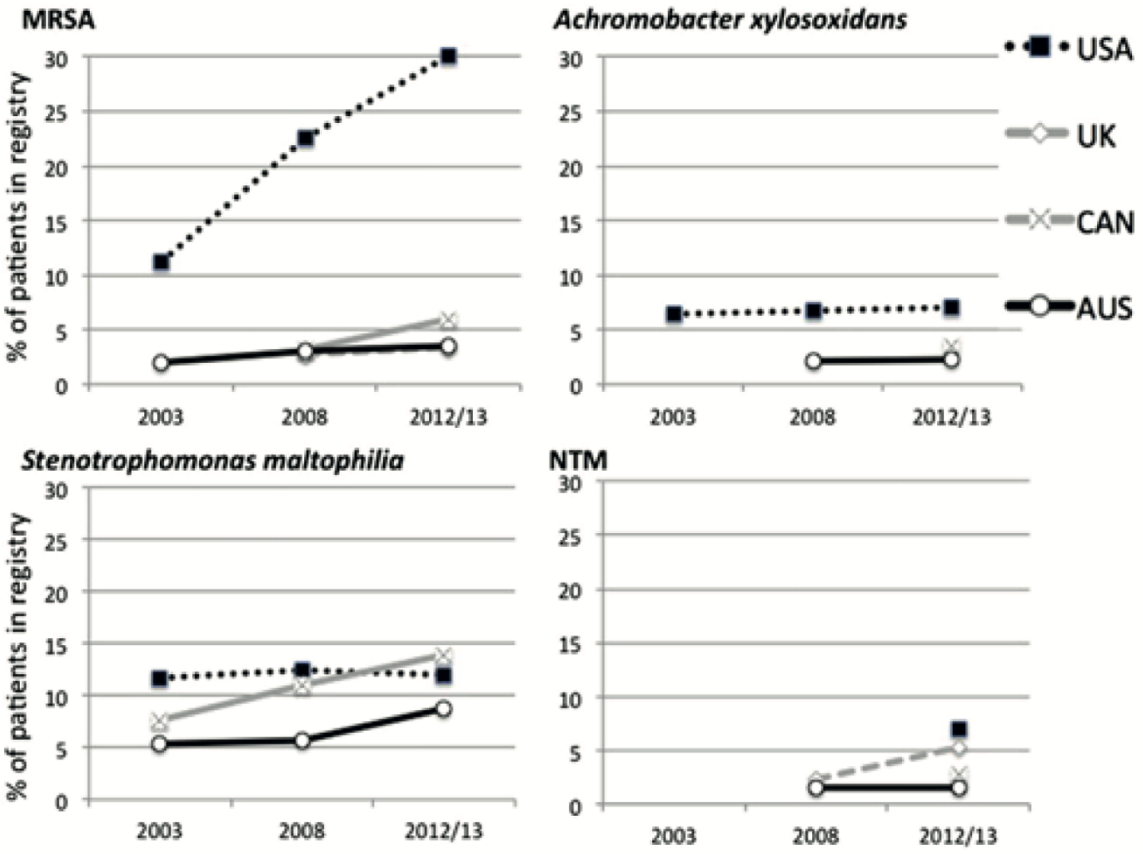


Figure 4: Pathogens increasing in prevalence in the CF community

Prevalence graphs showing the increasing prevalence of four emerging CF pathogens in four countries. Nontuberculous Mycobacteria (NTM) refers to all pathogenic mycobacteria that do not form part of the *M. tuberculosis* complex, including the *Mycobacterium abscessus* species complex (MABSC). Figure originally published by Parkins and Floto 2015 (38).

Several factors could be responsible for this changing epidemiology. For example, better species identification techniques have shed light on pathogens that may have originally gone unrecognized as causing a significant burden of infection in CF (38). Also, the increased life expectancy of people with CF results in them being exposed to more pathogens with their lungs in a more vulnerable condition for longer. Finally, novel treatments, whilst having an overall beneficial impact on disease outcome, can have negative consequences. This final factor has been hypothesized to be a reason for why the prevalence of NTMs has increased in the CF community as the use of long-term inhaled azithromycin therapy increases clearance of *P. aeruginosa* but potentially inhibits the innate immune system's ability to phagocytose NTMs (52). Amongst the NTMs is the *Mycobacterium abscessus* species complex (MABSC); a group of highly antibiotic resistant organisms, which are now isolated

from between 6.6% and 32.7% of people with CF and where the chance of treatment failure is as high as 50% (53-55).

Consequently, novel antibiotics and novel drug targets are urgently needed to treat the organisms of the MABSC. In order to develop these a greater understanding of how this group of organisms has evolved and adapted to become significant CF pathogens, as well as an increasingly common nosocomial pathogens associated with both pulmonary and skin and soft tissue infections (SSTIs), is required. This is the focus of this thesis.

1.2 The *Mycobacterium abscessus* species complex

1.2.1 Taxonomy

The genus Mycobacterium is part of the family Mycobacteriaceae, suborder Corynebacteriaceae, order Actinomycetales and phylum Actinobacteria (56). The phylogeny of the Mycobacterium genus broadly splits into two clades, one consists of rapidly growing mycobacteria (RGM), which are defined as those that produce colonies in 7 days and includes organisms such as the relatively rare human pathogens *Mycobacterium fortuitum* and *Mycobacterium smegmatis*, the other consists of slow growing mycobacteria (SGM), including the majority of the most notorious pathogens encompassed within this genus, such as *Mycobacterium tuberculosis*, *Mycobacterium ulcerans*, *Mycobacterium leprae* and the *Mycobacterium avium* complex (MAC) (Figure 5) (57, 58). NTM are all Mycobacteria that are not part of the *Mycobacterium tuberculosis* complex or *Mycobacterium leprae*. Amongst the rapidly growing NTMs and closely related to *M. chelonae* and *M. immunogenum* is the MABSC (
Figure 5).

The MABSC consists of three subspecies, *Mycobacterium abscessus* subspecies *abscessus* (*M. a. abscessus*), *Mycobacterium abscessus* subspecies *bolletii* (*M. a. bolletii*) and *Mycobacterium abscessus* subspecies *massiliense* (*M. a. massiliense*) (59). This classification has undergone several changes and is still debated amongst the scientific community. The differing antibiotic susceptibilities of the subspecies means their classification is of great clinical significance.

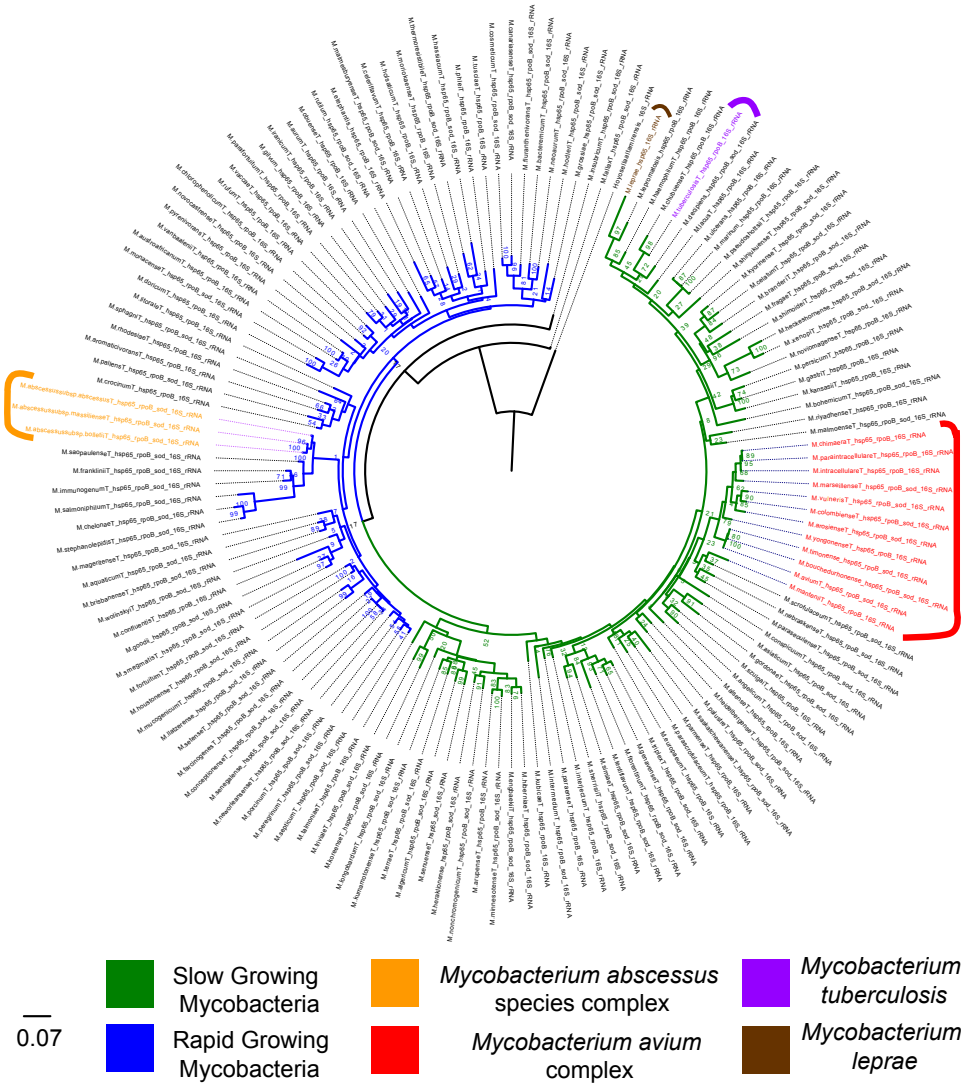


Figure 5: Phylogeny of the Mycobacterium genus

Multi locus sequence typing (MLST) maximum likelihood phylogeny of housekeeping genes *rpoB*, *sod*, 16s rRNA and *hsp65*. The partial sequences of housekeeping genes, 16s rRNA, *hsp65*, *rpoB* and *sod* that were used by Devulder et al. (2005) were used to query all the sequences used by Tortoli et al. (2017) in the most recent Mycobacterium genus phylogeny (57, 58). The partial sequences were extracted for the 144 sequences using nucleotide BLAST (version 2.7.0) and aligned using MUSCLE (v. 3.8.31) (60). A maximum likelihood phylogeny was inferred from the alignment of genes using RAxML (v. 8.2.8) (61). 100 bootstrap replicates were performed.

The first report in the literature of these organisms was in 1953, when atypical acid-fast staining bacilli were recovered from a deep knee abscess of a 63 year old women and subsequently classified as *Mycobacterium abscessus sp. nov* (62). In 1972 *M. abscessus* was re-classified as a subspecies of *M. chelonae* on the basis of overlapping biochemical

and phenotypic characteristics (63). It was reinstated as its own species in 1992 on the basis of DNA-DNA hybridization (DDH), which showed it shared less than 70% relatedness, the species cutoff, to *M. chelonae*, its ability to grow in the presence of 5% sodium chloride (NaCL) and inability to utilize citrate as a sole carbon source (64). The first reports of *M. massiliense* and *M. bolletii* occurred in 2004 and 2006 respectively, although *M. massiliense* was not acknowledged as a novel species officially until 2006 (65-67). On the basis of 16s rRNA and *rpoB* sequence similarity both species were shown to be closely related to *M. abscessus,* but it was not until 2009 that it was suggested that *M. massiliense*, *M. bolletii* and *M. abscessus*, indistinguishable with biochemical and phenotypic tests, were genetically related enough to form a single species (68). Initially, this was proposed to consist of two subspecies: *M. a. abscessus* and *M. a. massiliense* (which encompassed all isolates previously classified as *M. massiliense* and *M. bolletii*) (68). However, this was subsequently changed from *M. a. massiliense* to *M. a. bolletii* to follow classification rules (69).

Evidence from whole genome sequencing (WGS) very quickly emerged which contradicted this classification. Phylogenetic analysis showed that isolates from *M. abscessus*, *M. bolletii* and *M. massiliense* formed three monophyletic clades, significantly diverged from one another (70-73). Average Nucleotide Identity (ANI) analysis showed that representatives from the three species were above the species cutoff boundary of 95-96% ANI (74-76), but equally different from each other, suggesting that a subspecies relationship was appropriate and agreeing with previous DDH analysis (58, 68, 71, 73). Furthermore, the differing susceptibility to macrolide antibiotics, due to the presence and absence of a full length erythromycin ribosomal methyltransferase (*erm(41))* gene encoded by the two species, *M. bolletii* and *M. massiliense,* combined into subspecies *M. a. bolletii* by Leão et al. (2011), makes it essential that these subspecies are distinguished from one another (69, 77). This taxonomic change to show that MABSC consists of three subspecies has recently been proposed (59). Although, there is now debate within the scientific community as to whether this should be upgraded to a species level separation (58, 78). For the purpose of this thesis, Tortoli et al's (2016) description of the MABSC, where the MABSC consists of three subspecies, will be used.

1.2.2 Physiology

M. a. abscessus, *M. a. bolletii* and *M. a. massiliense* are Gram positive, acid-fast staining, non-motile, non-spore forming, rod shaped, obligately aerobic bacilli (67). The *M. a. abscessus* type strain ATCC19977, when grown on egg medium, produces after 7 days intermediary rough-smooth, white-grayish, non-photochromogenic rods 1.0-2.5µm in length,

0.5µm wide and has the ability to grow at 28℃ and 37℃ but not at 42℃ (62, 64). The *M. a. bolletii* type strain BD, originally isolated from a bronchial aspirate, grows in 2-5 days at temperatures between 24℃ and 37℃, producing non-pigmented colonies on 5% sheep blood agar, Middlebrook 7H10 agar and egg-based Löwenstein-Jensen (LJ) slants (66). Optimal growth occurs at 30℃ (66). Similarly to *M. a. bolletii*, the *M. a. massiliense* type strain CIP108297, originally isolated in Marseille from both the sputum and bronchoalveolar fluid of a patient with hemoptoic pneumonia, grows in 2-4 days on 5% sheep blood agar, Middlebrook 7H10 agar and egg-based LJ slants at temperatures between 24℃ and 37℃, but optimally at 30℃ (67). When grown on 5% sheep blood agar it produces non-photochromogenic colonies, with intermediary smooth-rough morphotypes, similar to those described for *M. a. abscessus* ATCC19977 (64, 67).

The physiological characteristics that are shared by all the subspecies of the MABSC include the inability to use glucose, fructose, citrate or oxalate as a sole carbon source, as well as the inability to synthesize Tween 80 hydrolase and nitrate reductase (68). All the MABSC subspecies are able to synthesize arylsulfatase but unable to uptake iron from an inorganic iron containing reagent (68). The MABSC along with *M. chelonae* can be differentiated from the *M. fortuitum* complex by this inability to reduce nitrate and uptake iron, whilst the MABSC can be differentiated from *M. chelonae* by its tolerance to 5% NaCl in LJ medium (although Adekambi et al. 2006 in their initial description of *M. bolletii* sp nov. state that it does not grow in these conditions (66)), its tolerance to 0.2% picrate and inability to use citrate as a sole carbon source.

The increased availability, affordability and resolution provided by genotypic and WGS species identification techniques means that phenotypic species identification techniques are becoming much less frequently used. There have been a large number of recent publications proposing techniques, such as Matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF), multilocus sequence typing (MLST), variable number tandem repeats (VNTR) and PCR based assays, for both identifying isolates belonging to the MABSC and distinguishing between the subspecies, however no gold standard MABSC identification has been settled upon by the scientific community. This has knock on consequences for the diagnosis and subsequent treatment plans for those who acquire a MABSC infection (79, 80).

1.2.3 Genetics

WGSs of the type strains of the three MABSC subspecies, *M. a. abscessus* ATCC19977, *M. a. bolletii* BD and *M. a. massiliense* CIP108297, provided the first insight into the general characteristics of MABSC genomes (71, 81, 82). MABSC organisms encode one circular chromosome, consisting of approximately 5Mbp. Similarly to all species in the actinobacteria genus the genomes of the MABSC have a high GC content of, on average, 64%. Interestingly, whilst *M. a. abscesuss* ATCC19977 has been found to encode a single ribosomal RNA operon, which is more typical of SGM as opposed to RGM, *M. a. bolletii* BD was reported, on the basis of coverage, to encode two (81-83). The number of ribosomal RNA operons encoded by *M. a. massiliense* CIP108297, was not discussed in the publication of the WGS. The genomes of the three MABSC type strains, *M. a. abscesuss* ATCC19977, *M. a. bolletii* BD and *M. a. massiliense* CIP108297 were annotated with 4,920, 4,923 and 4,828 CDSs[1] respectively, with pangenome analysis suggesting the MABSC has a core genome ranging between 3,354 and 3,947 CDSs (72, 84).

The contribution of mobile genetic elements to the genetic make-up of the MABSC type strains was not extensively investigated in the original publications, with the *M. a. abscessus* ATCC19977 genome the only type strain in which the mobile elements were discussed significantly (82). Five insertion sequences, three prophage like elements and a possible integrated plasmid, which encoded one copy of the insertion sequence, ISMAB1, were detected within the *M. a. abscessus* ATCC19977 chromosome (82, 85). A non-chromosomal based mobile genetic element in the form of an 18kb mercury resistance plasmid, pMAB23, with high similarity to pMM23 from *M. marinum* was also found to be encoded by the *M. a. abscessus* type strain. However, as more MABSC genomes have been sequenced the extent of the diversity present within the species complex has begun to become apparent and the possible contribution of evolutionary processes, such as gene gain and loss, the acquisition of mobile genetic elements, recombination and the transfer of plasmids, to the emergence of the MABSC as a more prevalent opportunistic pathogen has become of increasing interest (84, 86, 87).

Comparative genomic analyses of collections of MABSC isolates have begun to investigate the diversity of the MABSC. The first MABSC pangenome studies, which used 40 and 14 MABSC genomes respectively, found that the MABSC had an open pangenome, suggesting

[1] The WGS of *M. a. bolletii* BD and *M. a. massiliense* CIP108297 were reannotated in this study, with Prokka predicting these genomes to encode 4,903 CDSs and 4,821 respectively.

that there is further MABSC diversity to explore (72, 84). One explanation proposed for this high level of diversity has been the high number of phage that have been observed to have been gained and lost during the evolution of the species complex (88). As more MABSC WGS have been published there have also been several reports suggesting that MABSC genomes have a high level of plasticity and that large scale rearrangements and recombination events have occurred during their evolution (84, 89). However, few studies have attempted to characterize the level of recombination in the MABSC (86, 90). Sapriel et al's (2016) analysis, which was performed using housekeeping genes, showed that there was a high level of admixing between MABSC lineages and suggested that the MABSC was potentially using distributive conjugative transfer as a mechanism for diversification (86). They also observed that these highly admixed strains were more commonly associated with causing lung infections in people with CF (86). Similarly, Tan et al. (2017), using a WGS approach, found that recombination was potentially playing a more significant role in the evolution of the MABSC, and estimate that it has contributed to the diversification of the three subspecies to a greater extent than SNPs, as well as suggesting that both intra and inter-species recombination events had occurred (90).

MABSC phylogenetic analysis has also emphasized the level of diversity amongst the MABSC complex, however, it also showed that specific lineages of the MABSC appear to be more commonly isolated from people with CF as well as from MABSC nosocomial infections, which suggests that the genetic makeup of these particular lineages may enable them to be better able to cause infection in humans (Figure 6) (73, 89, 91). The genetic makeup of these more prevalent lineages has not been thoroughly investigated and there has been no attempt to examine whether convergence has occurred between the expanded lineages responsible for a significant proportion of MABSC infections from genetically distinct backgrounds. By investigating this it may be possible to identify genes beneficial to MABSC when infecting the human host and thus increase our understanding of the pathogenesis of these organisms.

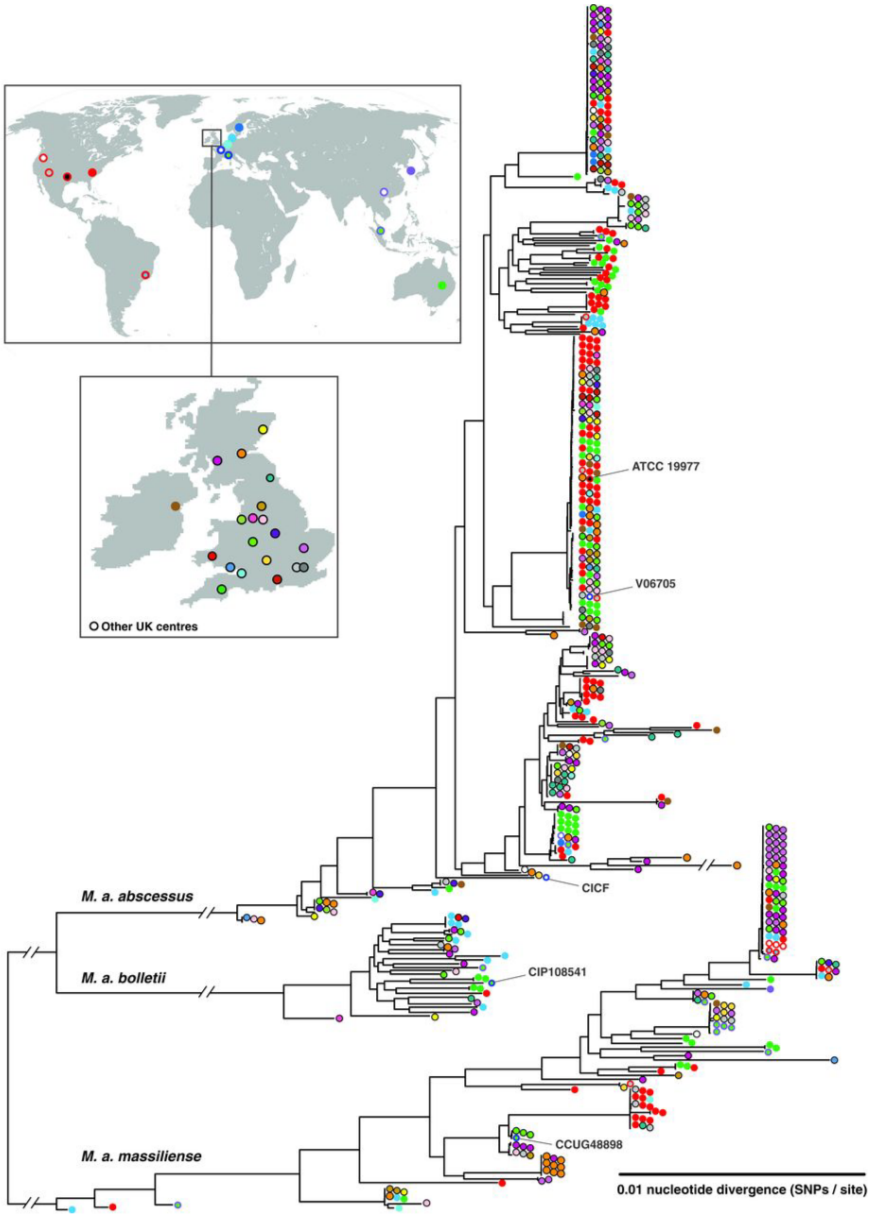


Figure 6: MABSC global population structure

Midpoint rooted maximum likelihood phylogenetic tree, constructed using RAxML. The MABSC global population structure shows that there is extensive genetic diversity with the MABSC but that there are also three dominant circulating clones. Phylogenetic tree constructed by myself, figure courtesy of by Andres Floto and published in (73).

1.2.4 Ecology

NTMs are environmental organisms found in water and soil worldwide (70, 92, 93). The environment has long been believed to be the main source of human infections caused by NTMs, although this has been debated since evidence of indirect person-to-person transmission of NTMs has been uncovered (70, 92, 93). It is notoriously difficult to isolate

NTMs from the environment and there are only a few examples of the causal agent of an human NTM infection being found in the environment (94-96).

MABSC isolates have most commonly been found in water as opposed to soil, although this may be due to soil culture being more susceptible to contamination thus making the isolation of NTM from soil more challenging (93). Isolates of MABSC have been recovered from both natural water systems, such as rivers, lakes and potentially even the ocean (as the long held belief that the salinity of oceans meant NTMs were not able to survive in this habitat was challenged by the isolation of several NTM species, including MABSC isolates, from wounds suffered during the Indian Ocean Tsunami of 2004), as well as man-made water systems, aided by the MABSC tolerance to chlorine, such as tap water, showers and hot tubs (97-100).

The ecological niches the MABSC can inhabit, particularly its presence in hospital tap water and on showerheads, places these microorganisms in habitats were they could come in contact with susceptible hosts (98, 100, 101). Furthermore, there is evidence that the MABSC grows within amoebae in its natural environment, which could unintentionally be preparing these organisms for life as intracellular pathogens, as amoebae are believed to reflect the macrophage environment, a theory that has also been proposed for *Legionella pneumophila* (65, 102). Additionally, the MABSC is able to form biofilms on pipe surfaces and on shower heads which suggested, and it has since been observed, that it would be capable of forming biofilms within the CF lung, a process that has been used by many CF pathogens, and provides protection against the host’s innate defense mechanisms and antibiotics (95, 103, 104).

Thus, the MABSC both inhabits environments that leave it perfectly positioned to cause opportunistic infections and already has innate features, evolved to aide the organism's survival in its natural environment, that could be beneficial to its survival in a human host.

1.3 The *Mycobacterium abscessus* species complex - the opportunistic pathogens

1.3.1 Epidemiology

The organisms of the MABSC are capable of causing pulmonary infections as well as SSTIs. Since the first report on NTM prevalence in the CF community in 1984, where the prevalence rate was estimated to be 1.3%, there has been a dramatic increase in prevalence with estimates as high as 32.7% having been reported in some CF populations (54, 105). However, the relative increase in isolation frequency has been greatest for the organisms that make up the MABSC (53, 106-109).

On the basis of the three largest NTM-positive culture prevalence studies conducted to date, with sample sizes of 986, 1216 and 1582 respectively, the prevalence rate is likely to be between 6.6 and 13.7% (53, 107, 108). However, prevalence estimates of NTM positive cultures isolated from CF sputum vary both by geography and the age range of the sampled population. Geographical prevalence differences as great as estimates of 0% and 28% and age group prevalence differences of 10% for children and 32.7% for adults respectively, have been observed (54, 110). The organisms of the MABSC are the most common NTMs isolated from CF sputum in Europe and Israel, whilst they are the second most common, after MAC, in North America (53, 111, 112).

This increase in NTM prevalence in CF and specifically the increase in the prevalence of the MABSC is not believed to be due to an increase in surveillance and/or improved detection of NTMs as firstly, multiple studies have shown that the prevalence is increasing annually whilst no change in surveillance frequency or identification techniques occurred (52), secondly, the relative frequency of the MABSC has increased in multiple countries (53, 106-109) and finally the increase in prevalence has also been reflected in an increase in NTM infections in non-immunocompromised individuals (113-116). However, the reasons for the increase in prevalence are unclear with increased environmental exposure, the increased use of antibiotics that create novel niches in the CF lung, the use of inhaled antibiotics which impair host immunity, or person to person transmission of NTM increasing its spread, all possibly contributing (52, 70, 73, 100, 109, 117-119).

Although less serious than MABSC pulmonary infections, particularly those in association with CF, the MABSC is also capable of causing SSTIs (120). Rare cases of ear infections,

ocular, central nervous system infections and bacteremia caused by the MABSC have also been reported (121-123). The global prevalence of extrapulmonary MABSC infections has not been thoroughly investigated, although reports on the burden of NTM extrapulmonary infections are beginning to be published and in several studies the MABSC has been found to be the most common cause of extrapulmonary NTM infections (124, 125).

MABSC SSTIs are generally acquired through wounds coming into direct contact with contaminated objects, for example tattoo needles or surgical instruments, or from a contaminated environmental source, such as spa water (126-128). Reports of MABSC wound infections after cosmetic surgery are becoming more frequent, particularly as medical tourism becomes more common (129).

Outbreaks of MABSC SSTIs have been reported, however, these are fairly rare events and generally only occur on a small scale and are usually acquired from a point source (127, 128). However, an exception to this is the large number post-surgical wound infections caused by the MABSC that have been seen in Brazil, where over 2000 cases have been recorded since 2004 (68). Molecular analysis of isolates obtained from Brazil has suggested that a single glutaraldehyde (GTA) tolerant lineage of *M. a. massiliense* is responsible (130, 131). Interestingly, phylogenetic analyses have suggested this lineage is closely related to lineages of the MABSC that have been associated with pulmonary infections outbreaks in CF centers, although the genetic distance between the Brazilian isolates in these studies did pose the question as to whether a single lineage was responsible for the upsurge in cases in Brazil (89, 91, 130).

The differing prevalence's of the three subspecies of the MABSC have not been thoroughly investigated, however, a study, investigating the global population structure of the MABSC in CF, using the largest collection of MABSC isolates to date suggests that *M. a. abscessus* and *M. a. massiliense* are responsible for the majority of the disease burden in CF, with far fewer *M. a. bolletii* isolates present in the collection (73). Whether this difference is due to differing environmental reservoirs, with *M. a. abscessus* and *M. a. massiliense* present in more human accessible environments, or differences in pathogenic potential is unclear.

1.3.2 Diagnosis

The increasing prevalence of NTM pulmonary infections led to the development of specific diagnosis guidelines. The fulfillment of clinical, radiological and microbiological criteria are required to diagnose a pulmonary infection, as a single positive culture indicating the

presence of an NTM is not necessarily indicative of infection. The current guidelines formulated by the American Thoracic Society (ATS) and Infectious Diseases Society of America (IDSA) require two clinical and radiological criteria to be met as well as one of several microbiological criteria (132). Firstly, the presentation of pulmonary symptoms with nodular or cavitary opacities on a chest radiograph, or a high-resolution CT scan showing multifocal bronchiectasis with multiple small nodules and secondly the exclusion of other diagnoses that could explain these symptoms must be satisfied (132). Microbiologically, one of the following has to be satisfied: two or more expectorated sputum samples must culture the same NTM species, an NTM positive culture should be cultured from a bronchoscopic lavage or wash or a transbronchial or other lung biopsy with mycobacterial histopathological features and positive culture of NTM combined with a sputum sample or bronchial wash that is culture positive for an NTM (132). For diagnosis of NTM pulmonary disease in a person with CF, the additional clinical criteria of worsening respiratory symptoms and/or declining pulmonary function tests that do not improve when the patient is treated with antibiotics specifically targeting conventional CF pathogens has been recommended (113).

It is critical to identify the NTM species responsible due to their differing antibiotic susceptibilities and in the case of the MASBC identification this is required down to the subspecies level for the same reason, although the continual taxonomic changes applied to the MABSC, as discussed above, have led to much confusion as to how to identify the MABSC subspecies. Antibiotic susceptibility testing is also recommended although there is often a discrepancy between *in vitro* and *in vivo* antibiotic activities (113). Therefore, even with successful identification of an MABSC pulmonary infection and the responsible subspecies, the current treatments available are often inadequate and the chance of successfully treating an MABSC infection remains widely regarded as unpredictable (133).

1.3.3 Treatment

The recommended treatment for a MABSC infection is with a macrolide, such as clarithromycin or azithromycin, and/or an aminoglycoside, such as amikacin, in combination with one or two parenterals, usually either imipenem, cefoxitin or tigecycline (113, 132, 134). The toxicity of many of these antibiotics often results in changes to treatment plans (55). In general the initial stage of treatment consists of 2-4 months of intensive treatment with an oral macrolide and intravenous antibiotics (cefoxitin, imipenem, tigecycline) followed by a continuation phase of treatment involving an oral macrolide, inhaled amikacin in combination with two to three additional antibiotics, either minocycline, clofazimine, moxifloxacin or

linezolid (113). Surgical excision of the infected tissue is also often required for a successful cure (55, 135).

Treatment outcomes have been shown to vary based on many factors including the subspecies involved, whether surgical resection of the infection site has been carried out, although this is not always possible, the severity of the disease and any underlying lung conditions and how well a patient can tolerate the currently available antibiotics (55, 136, 137). In fact, the lack of certainty in the outcome of treatment for MABSC pulmonary infections has led to an MABSC infection being seen as a contra-indication for lung transplant, currently the only cure for CF, in some CF centers, although this is under review (138-140).

The subspecies responsible for the infection can impact on the outcome due the differences in their innate resistance to Macrolide antibiotics. Macrolides are a key antibiotic used to treat MABSC infections, however, the presence of a full length *erm(41)* gene in *M. a. abscessus* and *M. a. bolletii* results in these subspecies having inducible macrolide resistance, unlike *M. a. massiliense,* and this believed to result in a better prognosis for patients infected with *M. a. massiliense (77, 141).* However, *M. a. abscessus* and *M. a. bolletii* isolates can become susceptible if they encode a cysteine amino acid (AA) at position 28 of the *erm(41)* gene as opposed to threonine at that position (142). Resistance to macrolides can also be acquired in all the MABSC subspecies by mutations in the 23s rRNA gene at positions A2048C and A2048G (142-146). This highlights the necessity for drug susceptibility testing as macrolides are the most effective drug against MABSC infections and there is currently no alternative with equivalent efficacy (133).

In the first study into the outcome of MABSC treatment in 1993, cure was achieved in just 8% of patients, with the majority of these patients also having surgical excision, whilst infection was fatal for 15% of patients, however this was carried out before the introduction of treatment with clarithromycin (135). More recent studies have highlighted the toxic nature of the treatment. Jarand et al. reported on the use of 16 different antibiotics in 42 different combinations and found that the majority of patients had to stop treatment with one drug, usually cefoxitin or amikacin, due to adverse side effects (55). Furthermore the fatality rate, given that this study was reporting outcomes of cases between 2001-2008, remained unchanged, at 15%, to that reported by Griffith et al. in 1993, whilst 50% of participants either remained culture positive or experienced a relapse (55). Further studies have reported similar rates of adverse reaction to the current treatment at 44% and rates of treatment

failure, either recurrence or death, in between 11% and 42% of participants and in some cases as high as 50% (136, 137, 147). There have also been reports of poor compliance to the recommended treatment guidelines (110, 148).

Whilst the toxicity of many of the antibiotics and compliance to those recommended are limiting factors in the treatment of MABSC infections, the main contributor to the limited treatment available and the poor outcomes associated with the MABSC is its highly antibiotic resistant nature, with the organisms of the MABSC widely regarded as the most antibiotic resistant NTM.

1.3.4 Antibiotic Resistance

The MABSC has innate, inducible and acquired antibiotic resistance mechanisms which result in this group of organisms being resistant to the majority of currently available antibiotics, including all the antibiotics used to treat *Mycobacterium tuberculosis* (149). The cell wall, characteristic of the Mycobacterium genus, provides a physical barrier to both antibiotics and biocides, preventing many classes of antibiotics, such as β-lactams, reaching their targets (149). The impermeability of the cell wall can be overcome however by porins which allow some antibiotics to enter the cytoplasm. However, once an antibiotic reaches the cytoplasm antibiotic resistance genes are expressed, which can upregulate efflux pumps resulting in the removal of the antibiotic from the cell or could be target modifying or antibiotic degrading enzymes (149).

Many antibiotic modifying or inactivating enzymes and target modifying enzymes are expressed by the MABSC (82). For example, the expression of the antibiotic modifying enzymes aminoglycoside 2-*N*-acetyltransferase and rifampicin ADP-ribosyltransferase could be responsible for the resistance of MABSC to aminoglycosides and rifampicin respectively, although in the case rifampicin this has yet to be confirmed (82). Macrolide resistance, as discussed previously, can be induced due to the presence of the *erm*(41) gene, a target modifying enzyme, which methylates the target site of macrolides on the 23S rRNA protein preventing the antibiotic from binding (77, 150). A *whiB7* family gene, MAB_3508c, is also encoded by *M. a. abscessus* ATCC19977, and similarly to its orthologs in *M. tuberculosis* and *M. smegmatis,* has been shown to control a regulon that is responsible for inducing resistance against multiple antibiotics (151, 152).

The MABSC is innately resistant to the first line *M. tuberculosis* drug ethambutol due to naturally occurring variants (I303Q and L304M) in the *embB* gene within the *embCAB* operon

which causes innate resistance as opposed to the accumulation of these variants after exposure to non-lethal levels in *M. tuberculosis* (153). Intrinsic resistance to fluoroquinolones in the MABSC is also due to naturally occurring variants in the DNA gyrase subunits *gyrA* (Ala-83) and *gyrB* (Arg-447 and Asn-464), with these variants also present and responsible for the intrinsic resistance in the MAC, *M. marinum* and *M. chelonae (154).*

Resistance to antibiotics has also been shown to be acquired by the MABSC after exposure to non-lethal levels of antibiotics. Both macrolide and aminoglycoside resistance can be acquired by mutations in the 23s rRNA (A2058G, A2058C, A2058T, A2059T, A2059C) and 16s rRNA (A1408G, T1406A, C1409A, G1491T) genes respectively (143, 155, 156).

To begin to combat the poor outcomes associated with MABSC infections, due to both these organisms' vast armory of defense mechanisms against currently available antibiotics and the toxicity of the antibiotics, a greater understanding of how the MABSC causes disease is needed.

1.4 Pathophysiology of the *Mycobacterium abscessus* species complex

1.4.1 Pathogenesis

The pathogenesis of the MABSC is poorly understood and much of what is believed to occur is based on what is known about the pathogenesis of *M. tuberculosis*. Pulmonary infections caused by the MABSC are either acquired by the inhalation or ingestion of contaminated aerosols, dust particles or water, whilst SSTIs are generally acquired through direct contact with a contaminated object (157). Similarly to *M. tuberculosis*, the organisms of the MABSC are intracellular pathogens and replicate within macrophages, although they are also capable of extracellular replication (158).

The MABSC is able to survive and replicate within macrophages through the blocking of phagosome and lysosome fusion, although the genes and pathways responsible are unclear (159). In some cases this is followed by the MABSC escaping the macrophage, although similarly exactly what genetic factors enable the bacterium to escape the phagosome have not been deciphered (158). These bacteria are subsequently able to infect more macrophages and can be released into the environment to potentially infect further individuals (158, 160). In other cases, the host responds to a MABSC infected macrophage

with the recruitment of lymphocytes and neutrophils to the site of the infected macrophage, resulting in the formation of a granuloma, an attempt by the host to contain the infection (158). This results in a chronic infection, with the bacterium entering a non-replicative state. A change in the host's immune status triggers the bacterium to begin replicating again and subsequently escape the granuloma and cause an acute infection (158).

The virulence factors enabling the survival and growth of MABSC in the CF lung have not been fully elucidated. Identifying the genes either essential to or involved in the pathogenicity of the MABSC would increase our understanding of how the MABSC is able to cause infection as well as potentially uncover novel drug targets.

1.4.2 Virulence factors

The first MASBC virulence factor to be thoroughly investigated was the distinct fate of MABSC isolates depending on whether they displayed a rough or smooth colony morphology (160). The differing morphotypes occur due to the disruption of genes involved in glycopeptidolipid (GPLs) synthesis and transport (MAB_4097c-MAB_4117c, MAB_0934-MAB_0939, MAB_4633, MAB_4437, MAB_4454c, and MAB_4459c) which results in the loss of GPLs from the cell surface (161, 162). The loss of GPLs expose the immunogenic lipids in the mycobacterium cell wall, which causes a greater pro-inflammatory cytokine response (163-165). Isolates with a rough morphotype have been shown to be better able to survive and grow intracellularly (166). Contrastingly, isolates displaying a smooth morphology, whilst being better able to form biofilms, are more likely to be controlled and cleared by macrophages (163, 167, 168). Spontaneous reversion from one morphotype to the other has been observed which suggests that both morphologies can be present at some point during an individual's infection (168).

As shown by the immunogenic nature of the rough MABSC morphotype, the cell wall, like in all pathogenic mycobacteria, is a key virulence factor in the MABSC. All mycobacterial cell walls consist of a plasma membrane, a core structure of peptidoglycan, arabinogalactan and mycolic acids, and an outer layer of glycolipids, polysaccharides, lipoglycans and proteins (169, 170). Further cell wall associated transport systems and lipids that have been shown to be involved in the pathogenesis of the MABSC, including the only two ESX type VII secretion systems encoded by the MABSC, ESX-3 and ESX-4 and *hadC*, a gene which functions in mycolic acid metabolism (171-173).

The WGS of *M. a. abscessus* ATCC19977 showed that *M. a. abscessus* encoded virulence factors typical of both mycobacterial pathogens and CF pathogens (82). Genes known to be associated with the virulence of other mycobacterial pathogens included phospholipase C, *mgtC*, mammalian cell entry (*mce*) genes, PE and PPE family genes, ESAT-6 family genes and lpqH like proteins. *Mce* genes have been linked with giving mycobacteria the ability to gain entry to mammalian cells and import host lipids for use as a carbon source, whilst PE and PPE genes, ESAT-6 family genes and lpqH family genes have been linked with roles in host pathogen interactions (174-177). Genes with functions that have been linked to the adaptation of other CF pathogens to the CF lung were also encoded. These included the phenylacetic acid degradation pathway, which has been shown to be essential for *B. cepacia* to be able to cause chronic infection (178), as well as genes associated with homogentisate catabolism (82). Homogentisate is a precursor in the production of the brown pigment pyomelanin, which has been linked, potentially by aiding iron acquisition, to the adaptation of *P. aeruginosa* to the CF lung (179).

RNA-seq analysis has suggested that in response to pressures known to be exerted within the CF environment, such as nutrient starvation, oxidative stresses and hypoxia, the MABSC changes the expression of pathways with similar functions to pathways known to have been involved in the adaptation of other pathogens, including *M. tuberculosis* and *P. aeruginosa* (180, 181). These pathways included the pyruvate dehydrogenase complex, fatty acid metabolism genes and the DosS/R regulon (180, 181). This suggested that the MABSC could be adapting to the host via routes that have been used by other intracellular pathogens (181). Similarly genes found to be differentially expressed by *M. a. abscessus* ATCC19977 when the organism was grown on synthetic CF medium (SCFM) showed that the MABSC upregulates pathways, such as those involved in amino acid metabolism, which have been shown to be used by other CF pathogens when adapting to the CF lung environment (181).

The detection of parallel evolution occurring as organisms evolve over time within patients has long been used to identify genes associated with the adaptation of the pathogens to their host. This method has been particularly commonly used to understand the adaption of CF pathogens to the CF lung. Through such analyses it has been shown that over time certain CF pathogens acquire mutations that cause a reduction in virulence as they adapt to cause chronic infection (40). The acquisition of hypermutator phenotypes has also been observed through this method (39, 182). This approach has been applied on a small scale to the MABSC with parallel evolution observed in the response regulator PhoR of the PhoPR two component system (TCS), as well as cell wall associated genes (183).

The detection of essential MABSC virulence factors has been and still is hampered by the difficulty in isolating the MABSC from the environment and thus it has not yet been possible to compare environmental and clinical MABSC isolates. However, the availability of larger WGS datasets means that it is now possible to examine virulence determinants with the benefit of their context in the genetic diversity displayed by disease causing MABSC lineages as opposed to examining virulence determinants encoded by a single reference genome, for which it unknown whether it present in the majority of disease causing MABSC lineages.

1.4.3 Transmission

A key change in the understanding of MABSC infections and a key driver in the potentially increasing prevalence of these infections, particularly with regards to people with CF, was the uncovering of evidence that indirect person-to-person transmission of MABSC isolates was possible (70, 119). However, our understanding of how MABSC has adapted to become transmissible is still very limited.

The first suggestion that person-to-person transmission of MABSC was feasible between people with CF was after an outbreak of *M. a. massiliense* in Seattle (119). The isolates from the five patients were indistinguishable via PFGE and epidemiological data showed the five patients overlapped in various hospital settings, suggesting contamination of the hospital environment by the index case, who began being treated in the clinic already infected with *M. a. massiliense*, resulting in indirect transmission to the subsequent cases (119).

A second outbreak, also involving a *M. a. massiliense*, occurred in Papworth hospital in the UK (70). 31 patients attending a CF clinic at Papworth hospital became infected with *M. a. massiliense* (70). WGS of the isolates recovered from these patients showed that there was greater genetic diversity within an individual patient than with other patients isolates, whilst phylogenetic analysis showed that the last common ancestor (LCA) of one patient's isolates was within the diversity of isolates from another patient, and furthermore isolates were found to be resistant to antibiotics that the patient they were recovered from had not been exposed to (70). The hospital environment was investigated and no environmental source for this lineage was found (70). Epidemiological information showed there were overlapping hospital stays between patients, whilst no interaction between patients had occurred outside the hospital (70). Taken together all these factors left indirect person-to-person transmission as the most probable explanation.

However, indirect person-to-person transmission of the MABSC remains a contested idea within the scientific community, with investigations in other hospitals not revealing evidence of transmission and sibling pairs, who have often been found to be infected with the same lineage of other CF pathogens, have been found to harbour distinct MABSC lineages (111, 184, 185). On the other hand, person-to-person transmission of the pathogen *Legionella pneumophila*, for which evidence first arose through WGS analysis and which was also contested amongst the scientific community, has recently been witnessed (186, 187).

Further evidence that transmission was potentially a significant contributor to the population dynamics of the MABSC infections in CF was uncovered after phylogenetic analysis of over 1000 MABSC isolates from around the world (73). This revealed the presence of three lineages, two within *M. a. abscessus* and one within *M. a. massiliense* that were responsible for over 50% of infection in CF and suggested that these lineages, which had potentially adapted to gain a selective advantage in the CF lung, were circulating globally amongst the CF community (73). The global distribution of these lineages could either be explained by these lineages being widely distributed in the environment, although their estimated recent emergence doesn't fit with this scenario, or the spread of these lineages via transmission (73).

Therefore, there is evidence to suggest that indirect person-to-person transmission is a significant factor in the emergence of the MABSC as a key CF pathogen, however, as of yet there has been no investigation into how these particular lineages of the MABSC have adapted to become transmissible.

1.5 Whole Genome Sequencing

1.5.1 Brief history of whole genome sequencing

Just 12 years after Watson and Crick discovered the structure of DNA in 1953, the first nucleotide sequence, that of an alanine tRNA from *Saccharomyces cerevisiae*, was sequenced (188, 189). In the 1960s sequencing of nucleotide sequences was reliant on RNAase enzymes and fractionation, with several short sequences successfully determined during the 1960s and early 1970s via these methods. However, it was in the 1970s that two techniques, Sanger sequencing and Maxam Gilbert sequencing, were developed that enabled the sequencing of longer nucleotide sequences and subsequently the sequencing of whole genomes (190, 191). Due to its greater efficiency and the fact it didn't require the use

of toxic chemicals, Sanger sequencing became the most commonly used sequencing method.

Sanger sequencing was developed by Frederick Sanger in 1977 (191). Sanger's method, also known as chain terminator sequencing, used labelled dideoxynucleotide triphosphates (ddNTPs) to terminate the extension of a DNA strand. In its early guise, four separate tubes were required each containing the template DNA, a DNA polymerase, a primer and deoxynucleotide triphosphates (dNTPs). ddNTPs representative of each base are added to individual tubes, with the DNA polymerase extending the primers till the addition of a ddNTP. After the newly synthesised strands were denatured from their templates, the DNA fragments from each tube were separated by fragment size using gel electrophoresis, with a lane representing each base. The DNA fragment bands were then visualised using autoradiography and the sequence read off. The introduction of fluorescent dye ddNTPs and capillary gel electrophoresis in later models, massively increased the efficiency and automation of Sanger sequencing (192).

The robustness and accuracy of Sanger sequencing meant it was the main sequencing method for many years until the development of next generation technologies which enabled the sequencing process to be massively parallelized. The first next generation sequencing (NGS) platform to be commercially produced was based upon the pyrosequencing method (193, 194). Pyrosequencing determines the sequence of a DNA fragment by utilising the inorganic pyrophosphate that is released upon the addition of a nucleotide by DNA polymerase as it extends a DNA strand. Through two further reactions the released pyrophosphate is turned into ATP which in turn acts as a cofactor for the oxidation of luciferin into oxyluciferin by luciferase, which results in light being emitted with the addition of base. The strength of the signal represents the number of bases added. By washing dNTPs representing each base over the plate separately, the sequence can be determined.

Life science technologies (which was subsequently purchased by Roche) developed the 454 sequencing platform which parallelized this method through the creation of a pyrosequencing assay whereby single stranded DNA fragments are ligated to individual beads before undergoing water-in-oil emulsion PCR, to coat the beads in clonal single stranded DNA fragments (195). The beads are then washed across a picolitre plate with a single bead filling a well. Pyrosequencing is then performed but with multiple sequences able to be sequenced in parallel. This was the method used to sequence one of the first human genomes (196).

However, the NGS technology originally developed by Solexa but purchased by Illumina has since become the most commonly used sequencing technology.

Illumina sequencing uses a sequencing by synthesis approach with reversible terminator chemistry (197). Single stranded DNA (ssDNA) fragments (~200-300bp) with adapters, indices and polymerase binding sites annealed to either end are washed across a flow cell with a lawn of oligonucleotides complementary to the adaptors attached. The DNA fragments bind their complementary oligonucleotides, DNA polymerase synthesizes a complementary strand and the original strands are washed away. This is to ensure the template strands used in the following steps are in the right orientation. These new strands now undergo bridge amplification to create clonal clusters. Clonal clusters are required so that the light emitted upon excitation of the fluorophore by a laser is of a great enough strength to be detected. The sequences representing the forward strand are washed away and sequencing by synthesis begins with the binding of the DNA polymerase to the read 1 primer binding site. Illumina sequencing uses modified fluorescently labelled dNTPs with reversible terminators. This allows the timing of the incorporation of each nucleotide to be controlled. The modified dNTPs are washed across the flow cell and the complementary base is added to the strand by the DNA polymerase, a laser allows the base to be imaged and the fluorophore is then washed away and the reversible terminator removed to allow the addition of the next base. This process is repeated for the desired number of cycles which corresponds to the read length. The sequencing product is then washed away. A DNA polymerase then binds the index 1 primer binding site and sequence by synthesis is performed. This enables samples to be multiplexed, with up to 96 bacterial genomes able to be sequenced in a single lane of a flow cell, thus dramatically increasing the cost effectiveness of the sequencing process. The index product is then washed away. The unannealed ends of the DNA strands are then bound to their complementary oligonucleotides attached to the floor of the flow cell. The primer binding site for the DNA polymerase to determine the sequence of the read two index is incorporated within the adapter, the DNA polymerase binds and sequencing by synthesis determines read 2’s index. The position of the cluster on the flow cell enables the illumina software to determine that this is the pair to the first sequence, as both pairs are sequenced in the same position. Bridge amplification is then performed again but afterwards the reverse strands of the original DNA fragment are washed away. This regenerates the clonal clusters. The final step is the sequencing by synthesis of the reverse strand using the forward strand as the template. By following this method paired end reads are generated. Paired end sequencing can be beneficial in resolving repetitive regions in genomes and for *de novo* assembly. However, it is also possible to do single end sequencing on the illumina platforms.

Whilst illumina sequencing is still the most cost effective and commonly used technologies globally, a third wave of sequencing technology, third generation sequencing, is now widely available and implemented. Single molecule real time (SMRT) sequencing is currently the most prominent of the third generation technologies. SMRT sequencing uses a DNA polymerase immobilized within a zeromode wave-guide (ZWG) of which there are hundreds in a single SMRT cell (198). The DNA polymerase binds to one of the hairpin adapters ligated to a double stranded DNA fragment of interest (198). Fluorescently labelled bases are then added to the SMRT cell and a video records the wavelength of light emitted as each base is added by the polymerase (198). As this method records the time taken between the addition of bases, referred to as the interpulse duration ratio, base modifications can also be detected through this method, enabling the researcher to investigate the methylome of the organism of interest (198). SMRT sequencing is also a long read technology, producing reads of 10-15kb in length, whilst the majority of NGS technologies are short read (including Illumina sequencing). Consequently, this platform is also useful for producing high quality reference genomes. Further third generation technologies such as the Oxford nanopore MinION and GridION platforms are also becoming more commonly used (199, 200). These technologies, as well as being long read, can sequence small genomes tin under an hour and have the potential to bring the ability to sequence DNA into any setting around the world or indeed into space (201, 202).

1.5.2 Microbiology in the genomics era

The first WGSs of bacteria, those of *Haemophilus influenzae* and *Mycoplasma genitalium*, were published in 1995 (203, 204). Since then, bacterial genomics has moved from comparisons between a few sequences, to exploring the population structure of species using thousands of isolates to tracking outbreaks using WGS in real time (201). Whilst the advancements in sequencing technology (discussed previously) since these first bacterial genomes were sequenced using Sanger sequencing have enabled these large datasets to be collected, a similar evolution in the tools and techniques available to investigate bacterial genomes has been required to take advantage of this new data.

The first bacterial genomic studies, along with aiding taxonomic classification, uncovered the virulence associated genes in notorious pathogens and emphasized the array of adaptive mechanisms, such as gene degradation, acquisition of mobile elements and recombination that bacteria use as they evolve to become best suited to an environment (83, 205, 206). The increase in the number of genomes available has led to the development of population

genomic approaches with the first pangenome analysis published in 2005 and the first WGS SNP based population structure in 2008 (207). The application of WGS in epidemiological studies has rapidly increased, with both local and long distance transmission events able to be resolved (208, 209). It is now possible to use Bayesian analysis to determine the mutation rate of a population and use this to infer the date of emergence of epidemic lineages and temporally track their spread (210). Tools to perform bacterial genome wide association studies (GWAS) have been developed and used to identify virulence and antibiotic resistance determinants in, amongst others, *Campylobacter jejuni* and *Streptococcus pneumoniae* (211-213).

It is now possible, through RNA-sequencing and SMRT sequencing, to examine genome-wide variation beyond changes to DNA. Through such analyses it has been possible to examine how changes in regulation and the methylome on a population level have contributed to the emergence of novel epidemic lineages, whilst these approaches also allow investigation into the regulatory responses of a bacterium to the environment within the host or after exposure to antibiotics (181, 214).

Thus, there is an ever increasing tool box available to investigate bacterial WGS datasets and it is with these tools that this thesis aims to investigate how the MABSC has evolved to become an increasingly prominent CF and nosocomial pathogen.

1.6 Thesis aims

The MABSC has emerged as a serious threat to people with CF as well as an increasingly common nosocomial pathogen causing both pulmonary and SSTIs. The poor treatment outcomes mean that novel treatments are desperately needed. The broad aim of this thesis is to increase our understanding of how the MABSC is adapting to become a human pathogen, through which targets for novel antibiotics may be discovered. This aim is addressed more specifically through the following questions:

1) What genetic changes drove the emergence of the MABSC lineages most commonly isolated from people with CF?

2) What genetic changes are driving the continuing expansion of the most prevalent MABSC lineages?

3) How is the MABSC adapting to the CF lung environment?

4) What can be learnt through whole genome sequencing of the largest known outbreak involving the MABSC - the epidemic of postsurgical wound infections in Brazil

2. Exploring the genetic determinants that drove the emergence of the three most prevalent *Mycobacterium abscessus* species complex lineages in the Cystic Fibrosis community

Statement of contribution: This project was designed and supervised by Julian Parkhill and Andres Floto. I performed all the bioinformatic analyses reported in this chapter. Dr. Sony Malhotra performed and interpreted the protein structure analysis performed in this study. Daniela Rodriguez-Rincon performed and interpreted the results of the virulence assays reported in this chapter, as well as extracting the DNA required for further sequencing. All the authors contributed to the interpretation of the results.

2.1 Introduction

The *Mycobacterium abscessus* species complex (MABSC) has emerged as a prominent threat to people with underlying lung conditions such as Cystic Fibrosis (CF). MABSC infections are the most common cause of non-tuberculous mycobacteria (NTM) lung infections in Europe and Israel, whilst they are the second behind *Mycobacterium avium* complex (MAC) infections in the USA (53, 111, 112). The increase in prevalence of MABSC infections concurrently in multiple locations suggests that the increase is genuine as opposed to being due to improved surveillance or NTM identification techniques (53, 106-109). It has also brought to attention the poor treatment available for MABSC infections, with the current MABSC treatment failing, due to either resistance or toxicity, in up to 50% of cases (55, 136, 137, 147). Furthermore, in some CF centres MABSC infections are seen as a contraindication to lung transplant, currently the only cure for CF, due to the high likelihood of re-infection (138). Consequently, with MABSC infections becoming more common worldwide, novel treatments are desperately needed.

With the number of MABSC infections increasing globally, a large dataset of over 1000 isolates from nine different countries, obtained mainly from people with CF, was sequenced in order to investigate the diversity of disease causing MABSC isolates (73). This dataset was supplemented by 29 publicly available WGS isolated from a further five countries. The resultant phylogeny revealed the presence of three large expanded lineages, two within *M. a. abscessus* and one within *M. a. massiliense,* as well as smaller clades of significantly densely clustered isolates and isolates from genetically distinct backgrounds (Figure 6) (73). Bayesian analysis showed that the three largest lineages, referred to as dominant circulating clones (DCCs), DCC1, DCC2 and DCC3, had emerged recently, in 1980, 1963 and 1972 respectively (73). The presence of isolates from different CF centres and different countries within each of the DCCs as well as many of the significantly clustered lineages showed that these lineages were widely disseminated (73). The pairwise SNP distance between the majority of isolates within these lineages was less than 20 SNPs and given that within individual patients the within host diversity ranged from 20 to 38 SNPs it suggested that the clustered lineages were spreading amongst the CF community via transmission (70, 73).

With 74% of the sequenced isolates falling within one of these densely clustered clades, it is evident that particular lineages are causing the majority of MABSC infections in people with CF and that transmission as opposed to environmental acquisition is the main route of infection (73). All of which is contrary to what had been observed in previous analyses which

had shown that people with MABSC infections tended to be infected with isolates from genetically distinct backgrounds indicative of acquisition from independent environmental sources (111, 184, 185). Why these recently emerged and widely disseminated lineages are responsible for the majority of infections in the CF community has begun to be investigated.

Correlating clinical metadata with the population structure showed that the clustered lineages were more commonly associated with a poor outcome, caused chronic infections more often and were more commonly resistant to aminoglycoside and macrolide antibiotics due to point mutations in either the 23s rRNA and 16s rRNA genes (73). Molecular phenotyping assays showed that the clustered lineages had significantly increased phagocytic uptake and survived for longer within macrophages and furthermore infection of severe combined immunodeficient (SCID) mice with clustered lineages led to significantly greater bacterial burden and granulomatous inflammation than infection of SCID mice with unclustered lineages (73). These results suggested that the clustered lineages were potentially dominating in the CF community due to having greater pathogenic potential, either due to increased virulence and/or transmissibility, than the unclustered lineages.

Whilst the molecular characteristics that explain why these lineages are thriving in the CF community have been described, the genetic determinants that are responsible for the success of the lineages in the CF community have not yet been investigated. Therefore, the aim of this project was to investigate the genetic changes that had occurred on the branches immediately before the clonal expansion of the clustered lineages, with specific focus on the three largest lineages, the DCCs. These changes may represent genes or variants that may have predisposed these lineages to be successful in the human host, or they may represent changes that occurred as the lineages started adapting to the CF lung environment, and were subsequently fixed before the major expansion of each lineage.

2.2. Methods

2.2.1 Mapping, variant calling and phylogenetic analysis

The 526 isolates[2] that make up the single isolate per patient MABSC dataset described in section 7.1.2.1 and 29 publicly available isolates were mapped to the *M. a. abscessus* ATCC19977 using BWA-MEM (v. 0.7.12) with the parameters described in section 7.3 (215). Variants were called using Samtools (v.1.2.1) and Bcftools (v.1.2.1) with the parameters described in section 7.3 (216). The variant sites were extracted from the alignment using SNP-sites (v.2.3.2) and a maximum likelihood phylogenetic tree was inferred from these sites using RAxML (v.v.8.2.8) with the parameters described in section 7.4 (61, 217).

In order to analyze the genetic changes occurring on the branches leading to the last common ancestors (LCA) of the DCCs, the SNPs were mapped back onto the phylogeny using the ACCTRAN parsimony algorithm applied via an in house script (developed by Simon Harris, see section 7.5 for further details) (218).

2.2.2 De novo assembly and annotation

De novo assemblies for the 526 isolates were constructed using Velvet (v.2.2.5) and Velvet optimizer (v.1.2) and annotated using the Prokka pipeline (219, 220). Further details about these methods are provided in section 7.6 and 7.7 respectively.

2.2.3 dN/dS analysis

To identify whether a change in selection pressure had occurred on the branches leading to the LCA of each of the DCCs, which could be indicative of adaptation to a novel environment, the ratio of nonsynonymous SNPs per nonsynonymous site to synonymous SNPs per synonymous sites (dN/dS) was calculated for each branch of the phylogeny using the Nei-Gojobori method and applied via an inhouse script (221). To examine how the dN/dS was changing over time, the dN/dS for each branch was plotted against time, using the number of synonymous SNPs accumulated on the branch as a proxy for time.

2.2.4 SNP density analysis

Due to the low number of SNPs accumulated by each gene on the branches leading to the DCCs, the dN/dS per gene could not be calculated. Therefore, in order to detect genes that

[2] Two isolates were originally mislabeled as belonging to different patients, after this analysis was performed these isolates were found to be from the same patient.

could have potentially predisposed the DCCs to thrive in the CF lung environment, SNP density analysis was performed to identify genes that had acquired a significantly different number of nonsynonymous SNPs on the branch leading to the LCA of the DCCs in comparison to the number accumulated on the branches representing the evolution of isolates that do not form part of any DCCs.

This was achieved by comparing, for each DCC, the number of nonsynonymous SNPs acquired by a gene on the branch leading to the DCC's LCA to the number of nonsynonymous SNPs acquired by the gene on all the branches representing the evolution of isolates that are not part of any of the DCCs. The number of nonsynonymous SNPs acquired by the gene on all the branches evolving independently of the DCCs was corrected for differing branch lengths, and thus the differing probability of the gene gaining a nonsynonymous SNP, by multiplying this number by the ratio of nonsynonymous SNP positions on the branch leading to the DCC to the number of nonsynonymous SNP positions on the branches evolving independently of the DCCs. A X^2 test was then performed to determine the genes that had acquired a significantly different number of nonsynonymous SNPs on the branch leading to the LCA of a DCC. P-values were corrected for multiple testing using the Holm method, with a p-value less than 0.05 seen as significant (222).

2.2.5 Pangenome analysis

To investigate whether the DCCs had gained an advantage over other MABSC lineages due to differences in their gene content and specifically whether the DCCs had gained the same genes or genes with similar functions, pangenome analysis was performed.

The MABSC pangenome was determined using Roary (<v.3.11.2), with a blastp percent identity threshold of 90% (223). In order to reduce the chance of including CDSs disrupted due to occurring over contig breaks or constructed using reads with low level contamination, only assemblies with less than 100 contigs and genome size less than 5.7Mbps were used as the input for Roary. This resulted in a final pangenome dataset of 512 assemblies (marked in appendix table 5.1). Initially, genes present in all isolates that made up a DCC and not present in isolates not associated with a DCC were investigated. However, this failed to identify accessory genes present in a large proportion of each DCC and therefore this criteria was adjusted to a gene of interest having to be: i) present in 90% of a DCC lineage and ii) not present in more than 10% of isolates that are not part of a DCC.

2.2.6 Candidate follow up analyses:

The Prokka annotations of the candidate genes were enhanced by searches against the Pfam (v.3.1.0) and InterPro (v.68) protein databases (224, 225). Functional understanding was also aided by assigning the candidate genes to their Clusters of Orthologous Groups (COGs). The genes encoded by *M. a. abscessus* ATCC19977 had previously been assigned to COG groups (162). The candidate genes identified through pangneome analysis were assigned to COGs, where possible, using EggNOG-mapper (v. 4.5.1) (226). PHASTER was used to identify if phage associated genes were amongst the candidates genes identified (227, 228). Where appropriate the Restriction Enzyme database (REBASE) was used to predict the methyltransferase type and the motif it potentially modified (229).

2.2.6.1 Reciprocal Blast to detect orthologous genes:

To enhance the functional understanding of the candidate genes the orthologous genes shared between MABSC isolates and *M. tuberculosis* H37Rv were identified. The orthologs between the *M. a. abscessus* ATCC19977 reference genome and *M. tuberculosis* H37Rv have been previously determined (230). A reciprocal blast approach, described in section 7.10, was used to determine the orthologous genes shared between the MABSC pangenome and *M. tuberculosis* H37Rv.

2.2.6.2 Functional enrichment and pathway analysis:

In order to investigate whether the DCCs were preadapted to the CF lung environment through changes in the same functional areas or through changes in the same pathways, gene ontology term (GO-term) enrichment and pathway analysis was performed. GO-terms had previously assigned to the CDSs encoded by the *M. a. abscessus* ATCC19977 reference genome (231). InterProScan was used to assign GO-terms to the 18,386 genes that were identified to be present in between 1-95% of the 512 MABSC isolates used in the pangenome analysis. The R package TopGO (v.2.20), was used to determine whether the candidates associated with the emergence of each of the DCCs were functionally enriched with particular GO-terms in either the molecular function (MF), biological processes (BP) or cellular component (CC) ontologies (232) using the parameters described in section 7.9. Pathway analysis was carried using the Blast2GO (v.4.1.9) interface as described in section 7.9 (233).

2.2.7 Follow up analysis on a methyltransferase potentially contributing to the success of the M. a. massiliense dominant circulating clone, DCC3

The pangenome analysis revealed the presence of a mobile element encoding a methyltransferase (dpnM) in all DCC3 isolates and just three isolates not associated with a DCC. To identify whether the mobile element encoding dpnM was present in isolates that were not included in the pangenome analysis, the raw reads of all the isolates used in this study were mapped to a reference of the mobile element extracted from *M. a. bolletii*, RHS37, using BWA-MEM following the methods described in section 7.3. The mobile element was identified as present if a depth of coverage of at least four reads on each strand was observed.

To investigate whether the presence of this mobile element had provided an advantage to the DCC3 lineage structural, molecular phenotyping and further bioinformatic analyses were carried out.

2.2.7.1 Structural modelling of dpnM

The following work was performed by Dr. Sony Malhotra. The protein sequence of the dpnM was compared against the RCSB PDB (Protein data bank) and found ID-2dpm chain A of *Streptococcus pneumoniae* to be a suitable template for modelling dpnM. BATON and FUGUEALI were used to align dpnM to the *Streptococcus pneumoniae* template, the modelling was then performed using MODELLER (234, 235).

2.2.7.2 Deletion of the mobile element encoding dpnM

The mobile element encoding dpnM was deleted from the DCC3 isolate BIR1049 using a modified mutagenesis by recombineering protocol for *M. abscessus* (236). This work was carried out by Daniela Rodriguez-Rincon. For full details of the method see Daniela Rodriguez-Rincon's PhD thesis (University of Cambridge, March 2018).

2.2.7.3 Complementation vectors

To confirm the phenotypic effect of the deletion of dpnM and the three mutations, D184A, Y187L and F42S, known to impact the function of proteins with structural homology to dpnM, four complementation vectors were constructed using the integrative vector pMV306-xylE as a backbone (237, 238). This work was carried out by Daniela Rodriguez-Rincon and for full details of the method see Daniela Rodriguez-Rincon's PhD thesis (University of Cambridge, March 2018).

2.2.7.4 Macrophage cell culture and phagocytic uptake and intracellular survival assays

To determine whether dpnM was playing a role in the increased phagocytic uptake or increased intracellular survival phenotypes associated with the clustered MABSC lineages Daniela Rodriguez-Rincon performed the assays described by Bryant et al. (2016) (73).

2.2.7.5 DNA extraction for single molecule real time (SMRT) sequencing

The following work was performed by Daniela Rodriguez-Rincon. Mycobacterial genomic DNA was extracted using a combination of bead beating and QIAmp mini kit (QIAGEN, UK) for the eight samples recorded in Table 1. DNA fragments above 10 kb, the optimal size for SMRT sequencing, were confirmed using a 0.9% agarose gel. For full details of the DNA extraction method see Daniela Rodriguez-Rincon's PhD thesis (University of Cambridge, March 2018).

Table 1: Isolates selected for SMRT sequencing to determine the motif recognized by dpnM

Isolate	Taxonomic position	dpnM
BIR1049	DCC3	present
SMRL154	DCC3 outlier	absent
DEN538	*M. a. massiliense*	Present
AUS856	*M. a. massiliense*	absent
RHS37	*M. a. bolletii*	present
DEN515	*M. a. bolletii*	absent
BIR1049 knock out	DCC3	absent
BIR1049 complemented WT	DCC3	present

2.2.7.6 SMRT sequencing and assembly:

The Pacific Biosciences RSII instrument was used to perform SMRT sequencing on 8 isolates. One SMRT-cell was used per isolate. Post sequencing analysis was performed using the SMRT-analysis.2.3.0 pipeline available via the SMRT-portal (239). The sequencing reads were assembled using HGAP v3 (240). This involves three steps. Firstly, pre-assembly which aims to produce long and accurate sequences. This is followed by the assembly of these high quality sequences into a draft genome and finally, the correction of the draft assembly by the PacBio RS_Resequencing protocol and Quiver (v1) (239). The approximate genome size parameter was set to 5Mbp (approximately the size of the reference genome *M. a. abscessus* ATCC19977) and the target coverage was set to 25.

2.2.7.7 Modification and motif analysis:

RS_Modification_and_Motif_Analysis.1 was run using the SMRT analysis software v2.3.0 embedded in the SMRT-portal (239). Briefly, this protocol uses SFilter to remove short reads and sequencing adapters. The filtered reads are then mapped to the assembly produced by HGAP using BlasR v1 (241). Kinetic analysis is then applied to the alignment of the reads to the reference enabling the identification of the modified bases by detecting bases where the interpulse duration ratio (IPDR) was significantly different from that of the *in silico* control (242). The modified motifs recognized by the methylases present in the genome were then identified using Motif Finder v1, with a minimum modification quality (MODQV) threshold of 30.

2.3 Results

2.3.1 No change in selection pressure on the branches leading to the LCA of the three DCCs

The MABSC global populations structure showed the presence of three large expanded lineages (the DCCs) consisting of isolates from multiple countries which suggested that there were epidemic lineages of MABSC circulating globally in the CF community (Figure 7) (73).

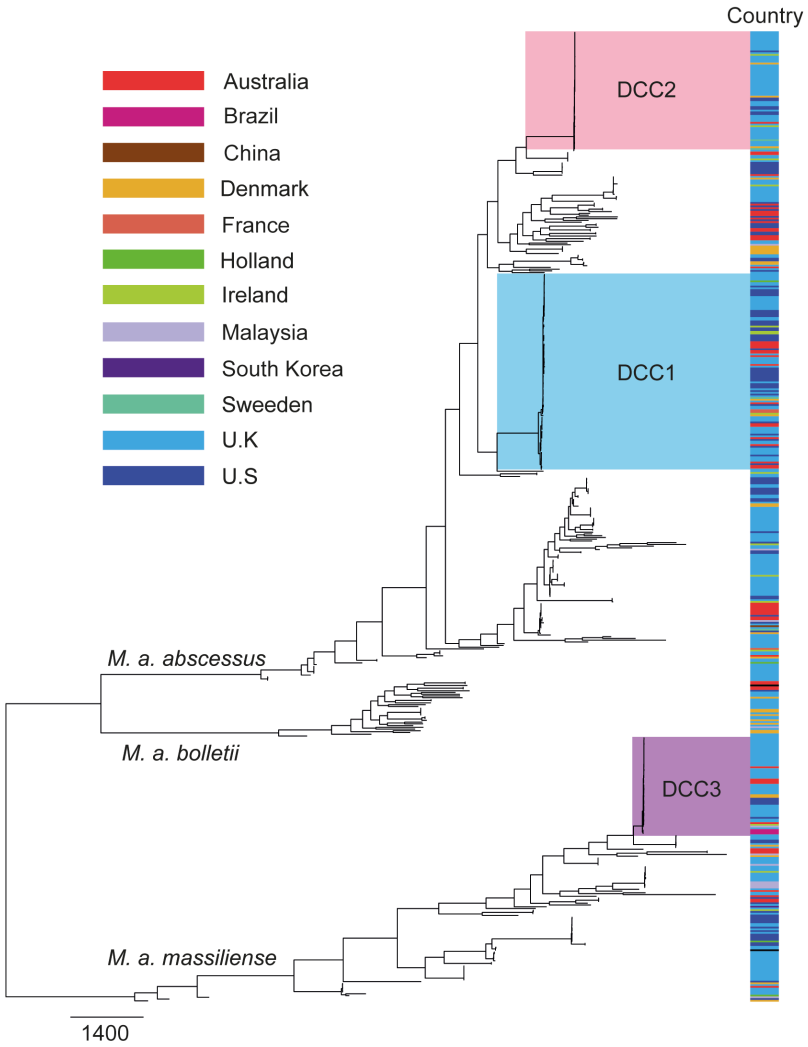


Figure 7: MABSC global population structure

Midpoint rooted Maximum likelihood phylogenetic tree of 555 MABSC isolates. This shows the clear differentiation of the three subspecies, *M. a. abscessus*, *M. a. bolletii* and *M. a. massiliense* that make up the MABSC. The recently expanded DCCs originally described by Bryant et al (2016) are marked in blue (DCC1), pink (DCC2) and purple (DCC3) respectively (73).

A change in selection pressure on the branches leading to the LCA of the DCCs would suggest that the lineages had adapted to a novel environment. Plotting the dN/dS values determined for each branch of the global population phylogeny against time, showed that on branches representing short time scales a wide range (0-7.52) of dN/dS values were observed (Figure 8). The vast majority of branches with high dN/dS values were at the tip of the tree, including many of the branches occurring after the clonal expansion of the three DCCs (points marked in blue (DCC1), pink (DCC2) and purple (DCC3) respectively in Figure 8). The small number of SNPs observed on these branches suggested that these were likely to be random fluctuations in dN/dS as opposed to a genuine signal of selection. Whilst the high dN/dS values at the tips of the tree were likely to have been effected by the fact that not enough time may have passed for purifying selection to have occurred.

Contrastingly, over longer time scales the dN/dS values converged to a value of approximately 0.1 (Figure 8), showing that the branches deeper within the phylogeny (grey points on Figure 8) or those leading to genetically diverse MABSC isolates (black points) were under strong purifying selection. The dN/dS values for the branches leading to DCC1, DCC2 and DCC3, 0.11, 0.11, and 0.12 respectively, fit with this trend, and indicated that no change in selection pressure had occurred on these branches (Figure 8 (inset)). This suggested that on the branches leading to the DCCs the majority of variants were accumulated when the lineages were evolving within their natural habitat, to which they were already adapted, and thus the majority of variants being accumulated were subject to purifying selection. However, amongst these variants there could be nonsynonymous variants, although not enough to cause a shift in dN/dS ratio, that potentially provided an advantage to the DCC lineages in the environment of the human host.

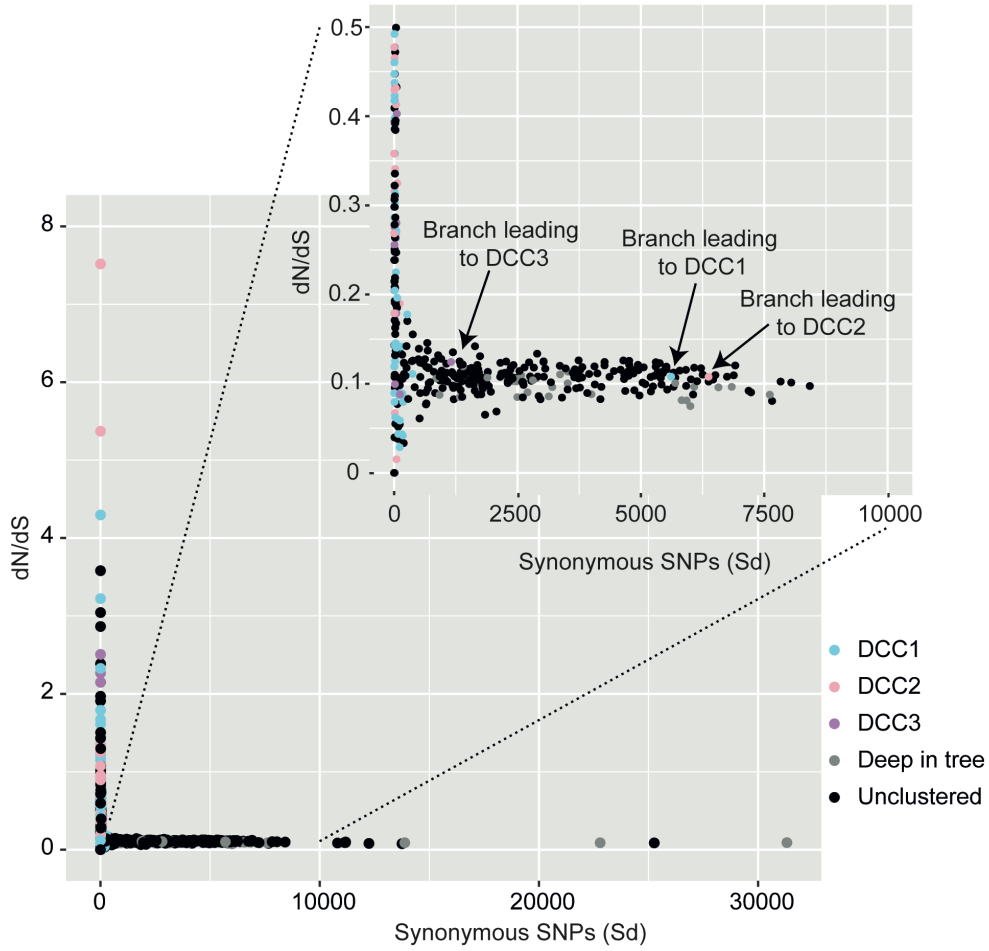


Figure 8: No change in selection pressure on the branches leading to LCA of the DCCs

dN/dS for each branch in the MABSC global population phylogeny plotted against the number of synonymous SNPs, which is used as a proxy for time. The branches leading to the LCA of DCC1 (blue), DCC2 (pink) and DCC3 (purple) are under strong purifying selection, with values of 0.11, 0.11 and 0.12 respectively.

To try and identify the genes containing these variants, the breakdown of synonymous and nonsynonymous SNPs per gene was determined. However, the low number of SNPs accumulated per gene on the branches leading to each of the DCCs meant that it was not possible to perform per gene dN/dS (appendix table 1.1). Therefore, SNP density analysis was used to try and detect genes that had accumulated a significantly different number of nonsynonymous SNPs on the branches leading to the LCA of the DCCs in comparison to the number accumulated by the gene during the evolution of isolates that did not form part of the DCCs.

2.3.2 SNP density analysis highlights functional areas where the DCCs may have undergone early adaptation to the CF lung environment

2.3.2.1 No evidence for changes in the same gene in the three DCCs before expansion

On the branches leading to the LCA of DCC1, DCC2 and DCC3, 23, 6 and 61 genes respectively were found to have accumulated a significantly different number of nonsynonymous SNPs (Figure 9, appendix tables 1.2, 1.3, 1.4). No gene accumulated a significantly different number of nonsynonymous SNPs on the branches leading to the LCA of multiple DCCs, which suggested that if the DCCs were preadapted for success in the human host, a single gene was not responsible. However, it did not rule out that genes with similar functions or participating in the same pathways had accumulated a significantly different SNP density on the branches leading to the LCA of multiple DCCs.

2.3.2.2 Genes assigned metabolism COGs are the most common amongst DCC1 candidates

Categorizing the 23 candidates identified as potentially associated with predisposing DCC1 to be more successful at causing infection in the human host into COGs, showed that genes with metabolism-associated functions were the most common, with 15/24 (65%) of the candidate genes assigned to metabolism-associated COGs. The most well represented COGS were amino acid transport and metabolism, and energy production and conversion, with four genes each. This was followed by three genes assigned to secondary metabolites biosynthesis, transport and catabolism, two genes assigned to the transport and metabolism COG and one gene assigned to carbohydrate transport and metabolism and nucleotide transport and metabolism respectively. The COGS not associated with metabolism that were also represented were transcription, with two genes and defense mechanisms with one gene. This suggested that DCC1 had potentially acquired advantageous changes in metabolic pathways in comparison to less prevalent MABSC lineages. Metabolic flexibility and adaptation is known to be key to the pathogenesis of *M. tuberculosis* and six of the genes annotated to metabolism associated COGs were found to be orthologous to genes in *M. tuberculosis* H37Rv (Figure 9, Appendix table 1.2) (243).

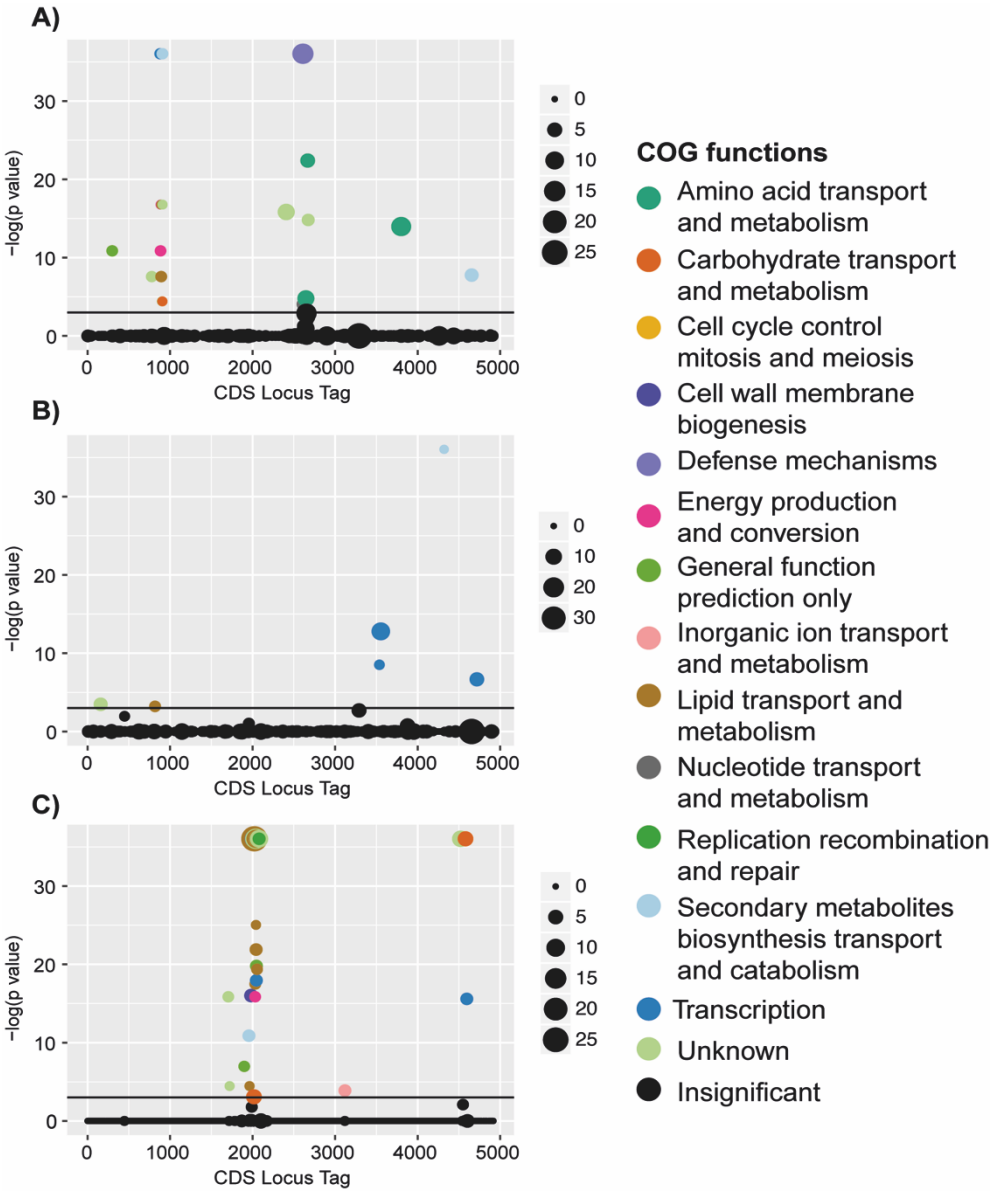


Figure 9: Genes with similar functions accumulated a significantly different number of nonsynonymous SNPs on the branches leading to the LCA of the DCCs

Manhattan plots showing the genes, colored by COG function, that gained a significantly different number of nonsynonymous SNPs on the branches leading to the LCA of A) DCC1 and B) DCC2 and C) DCC3. The black line marks the 0.05 significance threshold. The size of the points represent the number of nonsynonymous SNPs.

2.3.2.3 Changes to branched chain amino acid transport, biosynthesis and catabolism associated genes are prominent on the branch leading to the LCA of DCC1

MAB_0895 and MAB_0897 were found to be orthologous to Rv2495c (*bkdC/pdhC*) and Rv2497c (*bkdA/pdhA*) respectively (230). These genes form part of the branched chain alpha-keto acid dehydrogenase (BCKADH) complex, encoded by the *pdhABC* operon (244). This enzyme complex catabolizes the deaminated derivatives of branched chain amino acids (BCAA) into metabolites which are fed into the TCA cycle. Several genes encoded in close proximity to BCKADH enzyme complex genes also accumulated a significantly different number of nonsynonymous SNPs on the branch leading to the LCA of DCC1. These included MAB_0894c, annotated as a dihydrolipoamide dehydrogenase (*lpdA*), which could be encoding a key component of the BCKADH complex as well as two further enzyme complexes essential for virulence in *M. tuberculosis* (244). However, MAB_0894c was not found to be orthologous to *lpdA* in *M. tuberculosis* H37Rv (230). The presence of a further eight genes (MAB_0898-MAB_0918c) on the candidate list which are encoded in close proximity to the BCKADH complex genes in the *M. a. abscessus* ATCC19977 reference genome as well as the functions of these genes, which included a beta-ketoadipyl-CoA thiolase (MAB_0902) and Enoyl-CoA hydratase (MAB_0903) suggested that these genes could be involved in catalyzing reactions involving the products of the reactions carried out by the BCKADH complex and which result in the breakdown of BCAA intermediates into acetyl-CoA or succinyl-CoA which are subsequently fed into the TCA cycle (244).

Interestingly, the ATP-binding component of a BCAA ABC transporter operon, MAB_2622c, also accumulated a significantly different number of nonsynonymous SNPs on the branch leading to the LCA of DCC1. MAB_2622c (*livF*) forms the final gene in the *livJHMGF* operon. This operon has been shown to be important in the virulence of *Streptococcus pneumoniae,* whilst BCAA transport has also been shown to be important in the virulence of *Staphylococcus aureus* (245, 246). The ortholog to an enzyme involved in the biosynthesis of isoleucine, Rv1559 (*ilvA*), in *M. tuberculosis* H37Rv, MAB_2691, was also amongst the candidates. Rv1559, along with all the genes involved in the biosynthesis of BCAA, has been shown to be essential to the pathogenesis of *M. tuberculosis* H37Rv (247). Therefore there was evidence that changes had occurred in genes linked to the acquisition, biosynthesis and catabolism of BCAA on the branch leading to the LCA of DCC1.

Two further amino acid transport and metabolism associated genes were also found to have accumulated a significantly different number of nonsynonymous SNPs on the branch leading to DCC1 (Figure 9, appendix table 1.2). These included MAB_2699c, annotated as a

histidinol-phosphate aminotransferase (*hisC*), which was found to orthologous to Rv1600 (230). Histidine biosynthesis is essential for *M. tuberculosis* H37Rv survival within the host, whilst the histidine biosynthesis pathway has been shown to be up regulated by *M. a. abscessus* ATCC19977 in response to antibiotic exposure (181, 247).

2.3.2.4 Regulatory changes implicated in the emergence of DCC2

Six genes accumulated a significantly different number of nonsynonymous SNPs on the branch leading to the LCA of DCC2 (Figure 9, appendix table 1.3), three of which were annotated as regulators. The functions of the remaining three candidates were associated with lipid transport and metabolism, secondary metabolite biosynthesis, transport and catabolism with the function of the last being unclear.

The three regulators were all members of different regulatory families, with MAB_3565, a *tetR* family regulator, MAB_3582, a *gntR* family regulator and MAB_4754, an *araC* type family regulator (appendix table 1.3). The flanking genes of the regulators were examined in an attempt to predict which genes were under the control of the regulators (appendix table 1.5, 1.6, 1.7). MAB_3565 was flanked downstream by an alpha/hydrolase domain containing protein followed by MmpS and MmpL proteins. MmpL and MmpS play a role in transportation of lipids which form key constituents of the mycobacterial cell wall, which is in itself a key virulence factor in Mycobacterial pathogens (248). Upstream the functions of the genes were less clear, but they potentially have metabolism related functions, with a kynurenine formamidase/cyclase-like protein and thioesterase domain containing protein. MAB_3582 was flanked downstream by a thioesterase domain containing protein, followed by a *secA* gene and an AMP dependent ligase. *SecA* (MAB_3580) was found to be orthologous to *secA1* in *M. tuberculosis* H37Rv, which functions in protein transport, however, it does not have a known link to virulence in *M. tuberculosis* H37Rv, unlike its paralog *secA2* (249). Upstream MAB_3582 was flanked by a ribosome hibernation promoting factor, which was orthologous to Rv3421 and a hypothetical protein orthologous to Rv2342 (230). MAB_4754 was flanked downstream by an acyl transferase domain encoding gene, a gene encoding DUF222 as well as an HD domain encoding gene. Upstream a beta lactamase domain family protein was encoded followed by FAD domain encoding monooxygenase, a *tetR* family regulator and a acyl transferase domain encoding gene.

The non-regulatory candidates associated with the emergence of DCC2 included MAB_4353, a conserved hypothetical protein which was predicted to be a thioesterase family protein by Pfam and InterPro and was assigned to the secondary metabolite biosynthesis,

transport and catabolism COG (Figure 9, appendix table 1.3). This gene was found to be orthologous to Rv1532c, a conserved hypothetical protein encoded by *M. tuberculosis* H37Rv (230). A possible lipid metabolism associated gene, MAB_0827 and MAB_0162c, a gene of unknown function also accumulated a significantly different number of nonsynonymous SNPs on the branch leading to DCC2.

2.3.2.5 Metabolism COGs were the most common amongst the candidates associated with the preadaptation of DCC3

On the branch leading to the LCA of DCC3, 61 genes were found to have accumulated a significantly different number of nonsynonymous SNPs (Figure 9, appendix table 1.4). Figure 10 shows the distribution of candidate genes to each COG. Annotating the candidate genes with COG functions showed that 51% (31/61) of the candidates were associated with metabolism COGs, which suggested that DCC3 had, similarly to DCC1, been predisposed to be successful in the human host through changes in metabolic pathways. However, contrastingly to DCC1, genes associated with lipid transport and metabolism as opposed to amino acid transport and metabolism were the most commonly observed amongst candidate genes, with 29% of the candidates assigned to this COG (Figure 10). Transcription was the next most common COG with seven, whilst all the other COGs represented had less than five genes assigned (Figure 10).

2.3.2.6 A large proportion of the genes with significantly different SNP density on the branch leading to the LCA of DCC3 occurred in a known polymorphic region in MABSC genomes

A large cluster of the genes identified as possibly associated with the emergence of DCC3 as one of the most prevalent MABSC lineages consisted of a series of genes with nearly consecutive locus tags (MAB_2038-MAB_2101) (appendix table 1.4). The final three genes in this sequence, MAB_2099, MAB_2100 and MAB_2101, were annotated as a hypothetical cell division protein, a putative plasmid replication initiator protein and a recombinase.

Previously, this region of *M. a. abscessus* ATCC19977 has been hypothesised to encode a possible integrated plasmid and has been shown to contain the insertion sequence ISMAB1 (82, 85). The CDSs encoded within ISMAB1 (MAB_2086c-MAB_2088) were not found to have accumulated a significantly different number of nonsynonymous SNPs. However, 13 CDSs almost directly downstream of ISMAB1 (MAB_2070_MAB_2083), which were shown to have been spontaneously deleted in *M. abscessus* 390S in comparison to *M. abscessus* 390R, were found to have accumulated a significantly different number of nonsynonymous

SNPs (85). This suggested that many of the candidates associated with the emergence of DCC3 were encoded in a known polymorphic region of MABSC genomes (85).

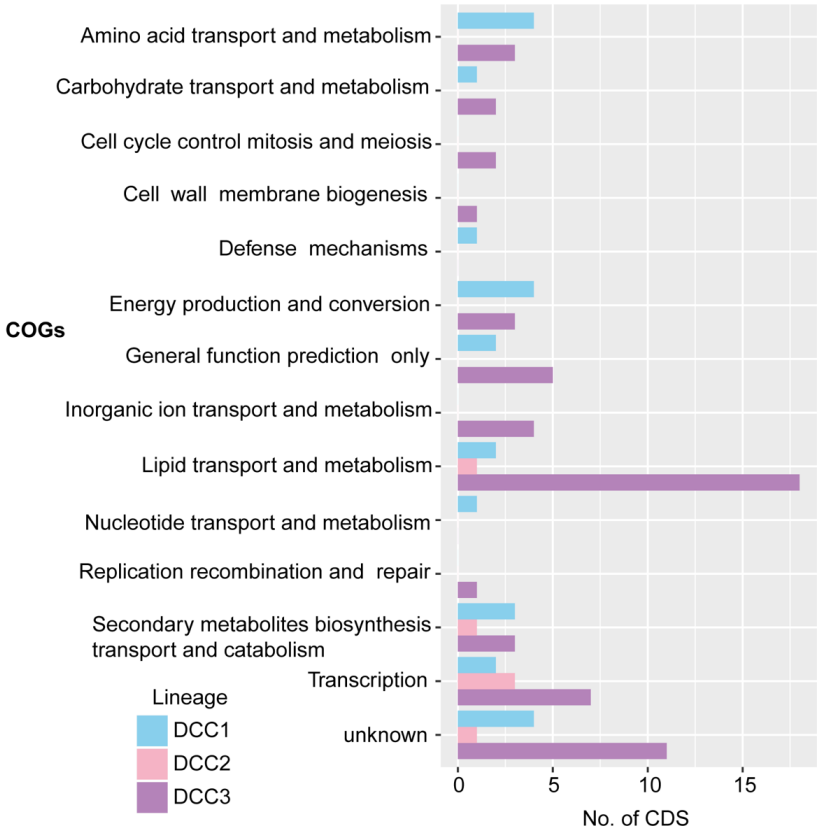


Figure 10: Distribution of COGs assigned to the SNP density candidates for each DCC

Breakdown of the number of candidate genes assigned to each COG category for DCC1, DCC2 and DCC3 respectively. For DCC1 the most common functions were amino acid transport and metabolism and energy production and conversion. For DCC2 transcription was the common function and for DCC3 lipid transport and metabolism was the most common.

2.3.2.7 Changes in fatty acid metabolism-associated genes prior to the emergence of DCC3

The CDSs encoded by the polymorphic region and in the flanking regions included multiple *moaC* domain containing genes, an enoyl-CoA dehydratase, long-chain and medium chain fatty-acid-CoA ligases, an acid-CoA ligase, multiple acyl-CoA dehydrogenases, the alpha and beta subunits of an acetyl/propionyl carboxylase (MAB_2066/MAB_2067), a short chain dehydrogenase, multiple monooxygenases and a cytochrome p450 enzyme (Figure 9, appendix table 1.4). Homologs of enzymes with these functions are known to participate in the beta oxidation of fatty acids (83, 250). Beta oxidation of fatty acids is essential for *M.*

tuberculosis to be able to persist in the phagosome and the sequencing of the first genome of *M. tuberculosis* (H37Rv) showed that genes with roles in fatty acid degradation were over represented (83). Fatty acid degradation pathways have also been observed to be important in other CF pathogens, such as *P. aeruginosa* (251).

2.3.2.8 Changes in SNP density in orthologs of M. tuberculosis genes on the branch leading to DCC3

Amongst the candidate genes that were not encoded in close proximity to each other in the genome were five genes predicted to be orthologous to genes in *M. tuberculosis* H37Rv (Appendix table 1.4). Two of these genes, MAB_2004, a UDP-N-acetylmuramoylalanine-D-glutamate ligase, and MAB_2005, annotated as FtsW, a cell division protein, were orthologous to Rv2155c and Rv2154c respectively. In *M. tuberculosis* these enzymes catalyze steps in the biosynthesis of the *M. tuberculosis* H37Rv cell wall (252). MAB_1984, a phospholipid acyltransferase, which was assigned to lipid transport and metabolism COG, was found to be orthologous to Rv2182c and MAB_1919, a conserved hypothetical protein with an alpha/beta hydrolase domain, was found to be orthologous to Rv2223c and Rv2235c (230). A major facilitator superfamily (MFS) protein, MAB_4615, also accumulated a significant difference in SNP density on the branch leading to the LCA of DCC3. MAB_4615 was found to be orthologous to Rv2265c, a conserved integral membrane protein (230).

Through this analysis genes which had acquired a significantly different SNP density on the branches leading to the DCCs were identified and evidence was uncovered of functional areas, such as the metabolism of BCAA, regulatory changes and fatty acid metabolism, in which each of the DCCs had potentially undergone changes that could have resulted in them being better able to survive and thrive in the human host.

2.3.2.9 No functional areas were statistically significantly enriched in multiple DCCs

To investigate whether the DCCs were enriched in the same functional areas GO-term enrichment analysis was performed using TopGO. GO-terms had been previously assigned to each of the genes encoded by *M. a. abscessus* ATCC19977 reference genome (231). At least one GO-term was assigned to 4,030 of the 4,920 CDS encoded by the *M. a. abscessus* ATCC19977 reference genome (appendix table 1.8). At least one MF GO-term was assigned to 3,535 CDSs, 3,557 CDSs were assigned at least one BP GO-term and 2,642 CDSs were assigned at least one CC GO-term. No GO-terms were found to be enriched amongst the candidates identified through SNP density analysis as associated with the emergence of

each of the DCCs after the p-values had been corrected for multiple testing (Table 2, appendix tables 1.9-1.17).

Table 2: Summary of the GO-term enrichment analysis for the SNP density candidates

Ontology		No. of candidates assigned GO-term	No. of GO-terms investigated	No. significant GO-terms	
				P-values not corrected	P-values corrected
Molecular Function	DCC1 (n=23)	22	774	1	0
	DCC2 (n=6)	5	774	2	0
	DCC3 (n=61)	51	774	8	0
Biological processes	DCC1 (n=23)	22	1254	1	0
	DCC2 (n=6)	5	1254	0	0
	DCC3 (n=61)	49	1255	2	0
Cellular component	DCC1 (n=23)	14	190	0	0
	DCC2 (n=6)	3	190	0	0
	DCC3 (n=61)	35	190	0	0

Blast2GO was used to map the candidates identified through the SNP density analysis to KEGG pathways. Eight (34%) of the 23 candidates identified for DCC1, one of the six candidates identified for DCC2 and 16 (29%) of the 61 identified for DCC3 mapped to a combination of 61 different pathways (appendix table 1.18). Figure 11 shows that when the pathways were grouped into general functional categories, the distribution of functions was similar to that found when clustering the candidates by COG functions, with a high proportion of amino acid metabolism pathways identified amongst the DCC1 candidates and lipid metabolism pathways found to be one of the most common groups amongst the DCC3 candidates. However, it was not possible to draw conclusions from this analysis due the small number of candidate genes assigned to KEGG pathways and the fact that the majority of the candidates were mapped to multiple pathways which meant it was unclear which pathway the gene was participating in.

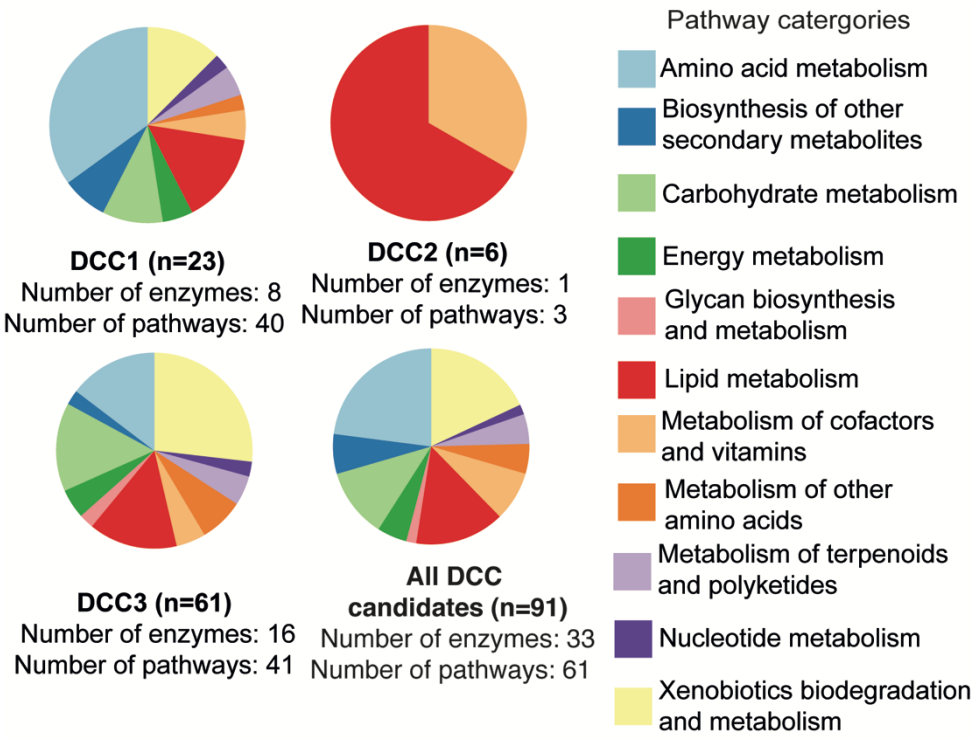


Figure 11: KEGG pathway analysis highlights similar functional areas to the COG analysis

Summary of the KEGG pathways that the SNP density candidate genes potentially participate in. 8/23 candidates for DCC1, 1/6 candidates for DCC2 and 16/61 candidates were mapped to a total of 61 pathways. Amino acid metabolism and lipid metabolism pathways are prominent in DCC1. Lipid metabolism pathways are prominent in DCC2 and lipid metabolism, amino acid metabolism and Xenobiotics biodegradation and metabolism are prominent in DCC3.

SNP density analysis highlighted changes that had occurred in the core genome that could have potentially preadapted the DCCs to be successful in the CF lung environment. Next, the MABSC pangenome was examined in order to investigate whether the acquisition of particular accessory genes had also contributed to the emergence of the DCCs.

2.3.3 The MABSC has an open pangenome:

A pangenome consisting of 35,994 gene clusters was identified by Roary from the 512 MABSC genomes analysed, with the total number of genes in the pangenome reduced to 35,678 when clusters with poor QC were removed (Table 3, appendix table 1.19). Table 3 summaries the number of core, soft-core, shell and cloud genes in the MABSC pangenome. The 316 gene clusters which failed QC were all cloud genes (Table 3)

Table 3: Summary of the MABSC pangenome

Number of isolates	All genes	Genes with poor QC removed
Core (>99%)	3,584	3,584
Soft core (95-99%)	271	271
Shell (15-95%)	2,010	2,010
Cloud (0-15%)	30,129	29,813
Total	35,994	35,678

Plotting the frequency of the 35,678 genes in each genome (Figure 12A), the total number of genes in the pangenome with the addition of each genome (Figure 12B) and the number of new genes introduced with the addition of each genome (Figure 12C) suggested that the MABSC has an open pangenome, with a large proportion of the genes present in only a few isolates, the number of genes in the pangenome continuing to increase and new genes still being observed even with the addition of the 512th genome. Previous analyses of the MABSC pangenome, albeit with smaller datasets, have also concluded that the MABSC to have an open pangenome (72, 84).

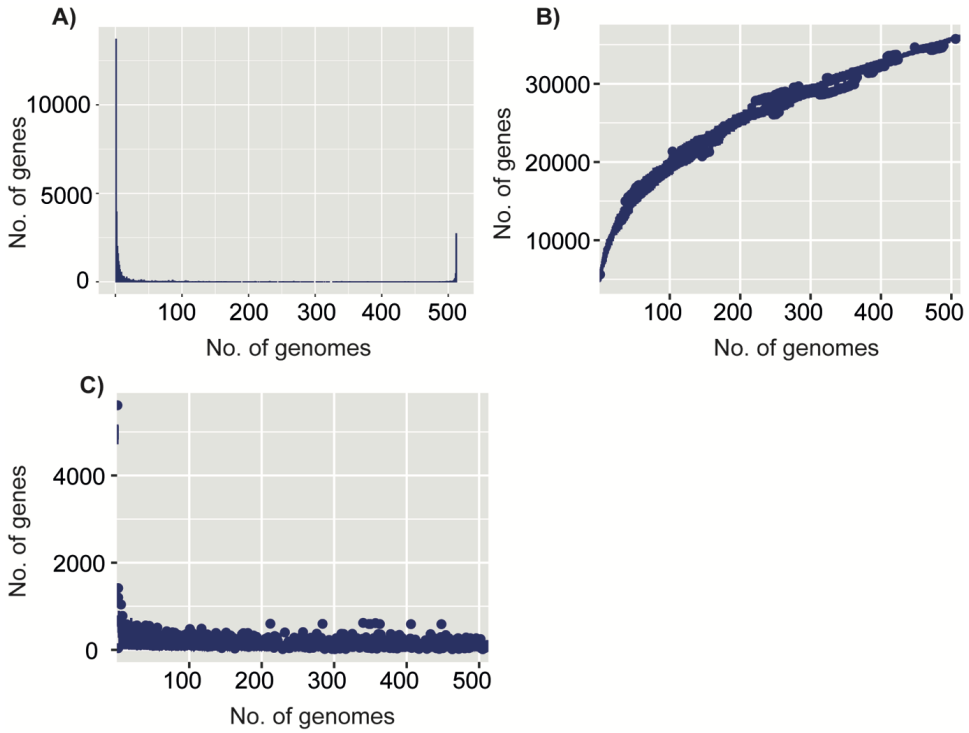


Figure 12: The MABSC has an open pangenome

Summary of the MABSC pangenome determined by Roary. A) Bar plot showing the frequency of each gene in the pangenome. B) Boxplots showing the number of genes in the pangenome as each genome is added. C) Boxplots showing the number of new genes observed with the addition of each genome.

An open pangenome is commonly observed in environmental bacteria, with the pangenomes of other environmental organisms, including those that can also cause opportunistic infections, also having been found to encode similarly open pangenomes (253). Furthermore, pangenome analysis has been able to shed light on the emergence of epidemic lineages of pathogenic bacteria, particularly if the emergence has been driven by the acquisition, via horizontal gene transfer (HGT), of genetic material (254). Consequently, the MABSC pangenome was analysed to investigate whether the three DCCs had acquired genes just prior to their clonal expansion that could have potentially given these lineages an advantage within the CF lung environment and particularly whether it was through the acquisition of the same genetic material that the three DCCs had been able to expand and become the most prevalent lineages in the CF community.

2.3.4 No evidence that the acquisition of the same gene or mobile element has driven the emergence of all three DCCs

Initially, the aim was to identify if any genes were present in all the DCC isolates and no isolates that were not part of the DCCs. Unique genes and genes encoded by isolates from differing DCCs were identified within the MABSC pangenome. DCC1 encoded 1,928 genes which were unique to the lineage along with 91 genes which were also encoded by isolates that formed part of other DCC lineages. DCC2 encoded 1,551 unique genes along with 73 genes also encoded by isolates from other DCCs and DCC3 encoded 501 unique genes and a further 26 genes that were also encoded by isolates from other DCCs. However, whilst genes encoded by all the DCCs and no other isolates not associated with the DCCs were observed, as well as genes uniquely encoded by two of the three DCCs, the vast majority of these genes were encoded by only a small proportion of isolates from each DCC (Figure 13).

The four genes encoded by at least one representative from each of the DCCs were annotated with potential virulence and antibiotic resistance associated functions, including an enhanced intracellular survival protein and an enzyme involved in cysteine biosynthesis (appendix table 1.20) (255-257). However, the low proportion of isolates from each of the DCCs encoding the four shared genes showed that the acquisition of these genes occurred after the LCA of the DCCs, and that these genes were potentially contributing to the ongoing expansion of the DCCs as opposed to providing the initial advantage that led to the clonal expansion of these lineages in the CF community (Figure 13).

A similar pattern was seen between the 62 genes unique to DCC1 and DCC2 isolates, the 25 genes unique to DCC1 and DCC3 isolates and the seven genes shared between DCC2 and DCC3 (Figure 13). On average only two DCC1 and DCC2 isolates each encoded one of the unique genes shared between DCC1 and DCC2, an average of two DCC1 and five DCC3 isolates encoded the genes shared by these lineages and an average of eight DCC2 and five DCC3 isolates encoded the shared genes between these lineages.

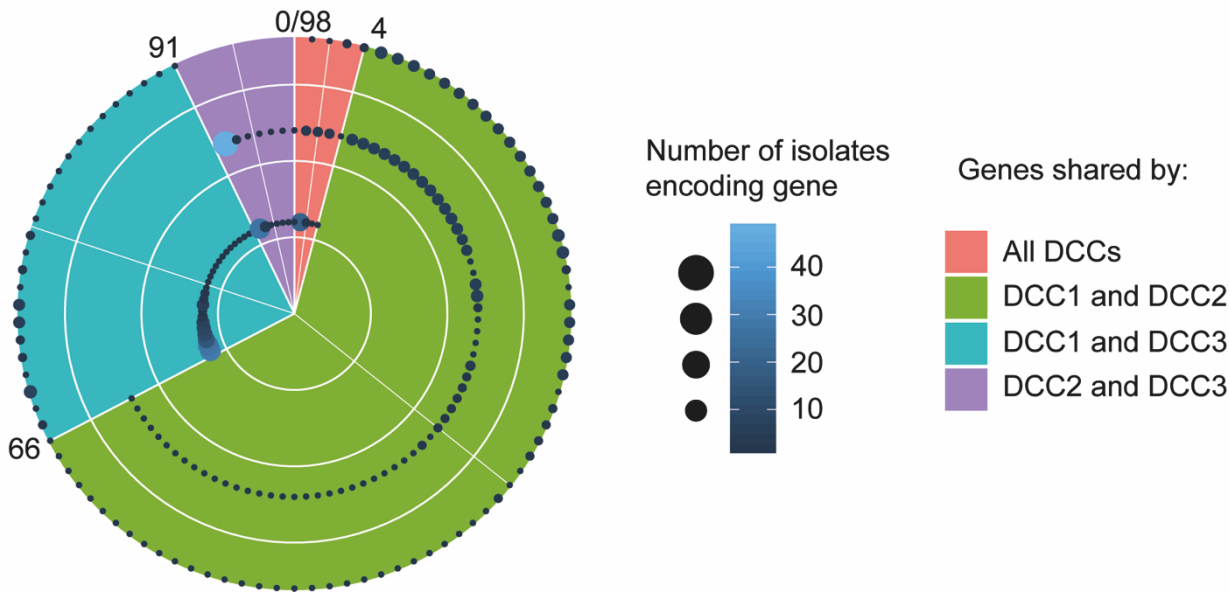


Figure 13: Genes unique to DCC lineages only shared by a small proportion of each DCC

Concentric circle plot with the outer circle representing DCC1, middle circle representing DCC2 and inner circle representing DCC3. Each point represents one of the 98 genes which are unique to DCC isolates and either present in a representative of all the DCCs (pink), representatives of DCC1 and DCC2 (green), representatives of DCC1 and DCC3 or representatives of DCC2 and DCC3. The size and color of the points indicate the number of isolates from each DCC which encode the gene.

The lack of shared genes by all DCC isolates suggested that the lineages were not predisposed to be more successful than other MABSC lineages in the CF environment through the acquisition of the same genes prior to their clonal expansion.

2.3.5 The majority of genes unique to individual DCC lineages were acquired after their LCA

The majority of genes found to be unique to individual DCCs were only encoded by a small proportion of each DCC's isolates (Table 4). Only in DCC1 were genes uniquely present in all the isolates of a single DCC detected, with 17 genes found to be uniquely encoded by all 105 DCC1 isolates. The consecutive nature of the locus tags and the fact that the same proportion of isolates gained the genes, suggested that these 17 genes represented the acquisition of two mobile elements; the first consisted of 10 genes (10208_3#20_03213-10208_3#20_03222), two of which encoded a resolvase and integrase respectively, and the second consisted of six genes (10208_3#20_03260-10208_3#20_03265). The final gene appeared to have been acquired independently.

Table 4: The proportion of each DCC that encoded the genes unique to each DCC

DCC		DCC1 (n=105)	DCC2 (n=63)	DCC3 (n=48)
No. of gene unique to DCC isolates		2,019	1,624	537
No. of genes present in n% of the DCC isolates with genes also present in other DCC isolates included but no genes present in non-DCC isolates included	100%	17	0	0
	90%-99%	2	31	0
	80%-89%	5	0	0
	70%-79%	3	1	0
	60%-69%	0	91	1
	50%-59%	2	13	34
	40%-49%	2	6	25
	30%-39%	2	5	1
	20%-29%	5	7	4
	10%-19%	61	151	14
	0-9%	1,920	1,319	458

No genes were present in all DCC2 isolates, however, 31 genes were present in 90-99% of the 63 DCC2 isolates (Table 4). Similarly to the DCC1 unique genes, many of them appeared to have been acquired together within three potential mobile elements. One consisted of three genes (10071_6#72_02556-10071_6#72_02559), the second consisted of 22 genes (10071_6#72_03928-10071_6#72_03957) and the third consisted of five genes (10071_6#72_03980-10071_6#72_03988). Given the close proximity in the genome of the second and third mobile elements it was possible that these were acquired together. Contrastingly, no genes uniquely encoded by DCC3 were present in greater than 69% of DCC3 isolates, which suggested that prior to the LCA of the DCC3 lineage, no acquisition of novel genetic material unique to the lineage had occurred.

The limited evidence of there being unique genes encoded by all the isolates which made up each of the DCC lineages and the lack of evidence of shared genes between a large proportion of all the DCC isolates suggested that the criteria applied to detect candidate genes acquired prior to the LCA of each DCC was too stringent. Consequently, in order to focus in on genes acquired by the majority of each of the DCC lineages (or that were acquired before the LCA and subsequently lost in some DCC members) whilst still selecting

for genes predominantly only present in DCC lineages, the selection criteria was relaxed to incorporate genes present in at least 90% of a DCC lineage isolates but not present in more than 10% of isolates not associated with a DCC.

2.3.6 Acquisition of gene clusters characterizes the early ancestral lineages of the DCCs

Adjusting the criteria for identifying the genes which potentially provided the initial advantage that drove the DCCs to become the most prevalent in the CF community, resulted in genes present in a large proportion of each of the DCCs individually being uncovered, as well as genes present in a large proportion of more than one DCC.

In total, 183, 217 and 119 genes were identified as potentially associated with the emergence of DCC1, DCC2 and DCC3 as the most prevalent MABSC lineages in the CF community (appendix table 1.21, 1.22, 1.23). No isolates that were not part of a DCC lineage shared exactly the same accessory genome, following this criteria, as any of the DCC lineages, suggesting that whilst some of the accessory gene content in each DCC lineage overlapped with both other DCCs as well as non-DCC isolates, at the point of the LCA of each DCC lineage, each had a unique combination of genes that could have contributed to the initial advantage these lineages had over other MABSC lineages in the CF lung environment (Figure 14). The basal lineage to DCC3, however, was shown to encode nearly all the candidates identified as present in a significant proportion of DCC3 isolates, with only 18 of the 119 DCC3 candidates not encoded by the basal lineage (Figure 14).

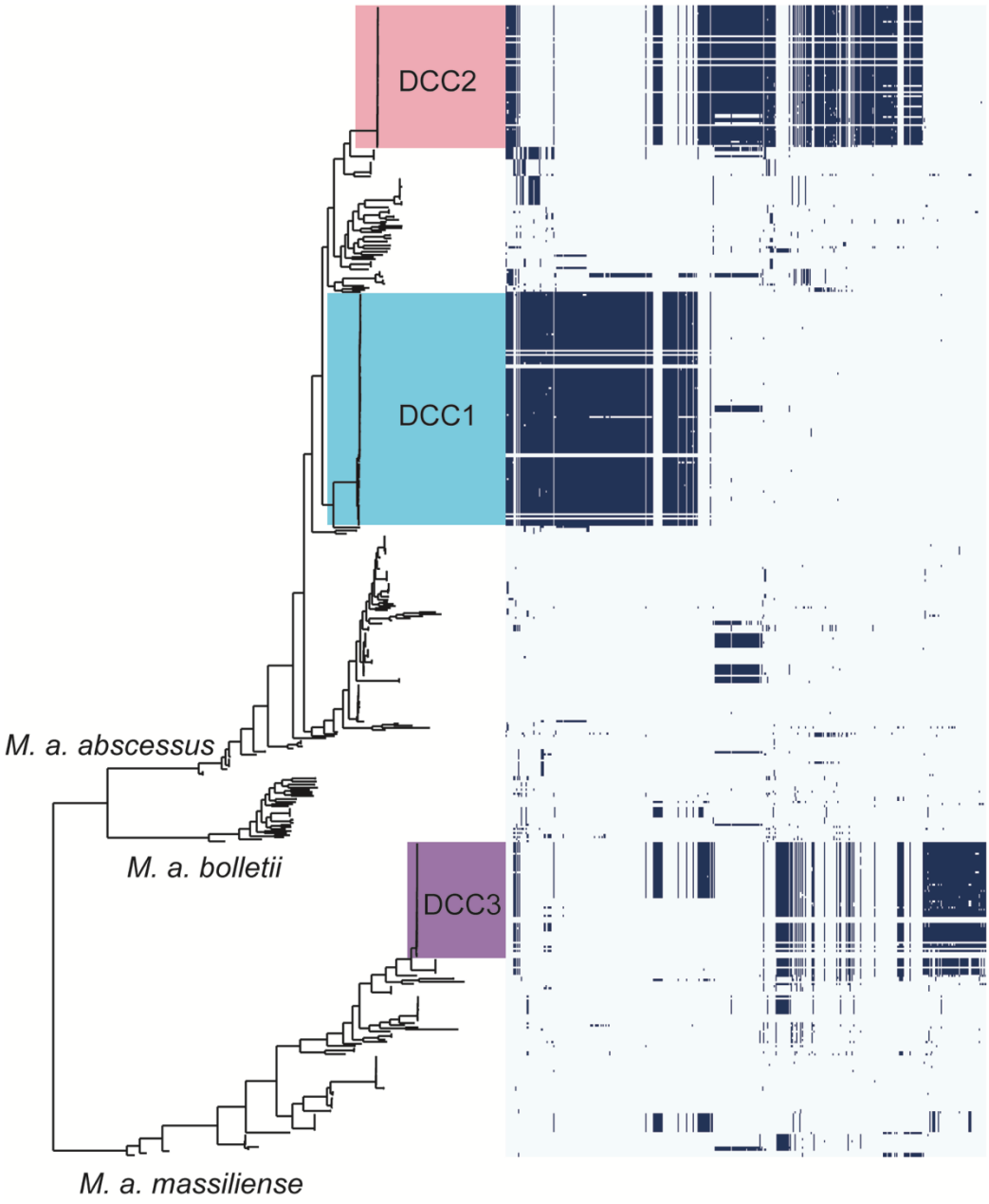


Figure 14: Each DCC encodes a unique array of genes potentially associated with virulence

The MABSC global population structure with the metadata representing the genes present in at least 90% of a DCCs isolates and less than 10% of isolates that do not form part of a DCC across the MABSC global population. Whilst there is overlapping gene content between DCC and non-DCC lineages, no other lineages encoded the same combination of accessory genes as each of the DCCs.

Within each candidate list, by using the Prokka predicted CDSs from the *de novo* assembly of a representative isolate from each DCC, many of the candidate genes were found to be encoded by consecutive locus tags[3] as well as in similar proportions of isolates (appendix table 1.21, 1.22, 1.23). Table 5 summarizes the number of candidate genes identified as associated with the emergence of each DCC and the number that were acquired with neighboring genes. The fact that many of the genes within each of the candidate lists were found to be encoded with neighboring genes, suggested that they were potentially acquired together, which in turn could imply that they function together.

Table 5: Summary of the number of genes in the DCCs accessory genomes acquired in clusters

	DCC1	DCC2	DCC3
Total number of genes	183	217	119
Total number of mobile elements (number of genes encoded by mobile elements)	11 (n=169, 92%)	19 (n=207, 95%)	14 (n=101, 85%)
Number of genes acquired independently	14	10	18

2.3.7 Overlapping accessory gene content with possible virulence functions between pairs of DCC lineages

Adjusting the criteria to select for genes present in 90% of a DCC lineage and less than 10% of isolates that did not form part of a DCC did not reveal any genes present in a significant proportion of all the DCC isolates. One gene, an efflux pump, was identified which was encoded by greater than 90% of DCC2 and DCC3 isolates and a single DCC1 isolate (Figure 16, 17). This was the only gene present in all the DCCs and was found in greater than 90% of two of the DCC lineages. However, there were examples of genes present in a large proportion of two of the DCCs but not present in the third.

[3] Empirical judgement was used to decide how many missing locus tags it was reasonable to allow – this ranged from 1-2 missing loci in a sequence of consecutive locus tags.

2.3.7.1 Ubiquinone biosynthesis associated operon present in both DCC1 and DCC2

Over 90% of DCC1 and DCC2 isolates each acquired an operon that encoded a *lysR* family regulator, followed by an L-carnitine dehydratase, a flavin protein decarboxylase (*ubiX*) and a 3-octaprenyl-4-hydroxybenzoate carboxylase (*ubiD*) (Figure 15,16). UbiX and UbiD catalyze the step in the biosynthesis of ubiquinone (coenzyme-Q) which results in the conversion of 3-octaprenyl-4-hydroxybenzoate to 3-octaprenylphenol (258). In *E.coli* ubiquinone biosynthesis has been shown to be involved in aerobic respiration, adaptation to oxidative stress and gene regulation (258).

A further four genes were also present in over 90% of both DCC1 and DCC2 isolates, with two encoded by consecutive locus tags (10208_3#20_00377-10208_3#20_00378). The functions of these genes were unclear, although 10208_3#20_00377 was found by Pfam to potentially be a transmembrane protein.

2.3.7.2 mce genes and oxidative stress related genes present in large proportion of DCC2 and small proportion of DCC1 isolates

There were also two, or potentially one given their presence in similar proportions of isolates, clusters of genes acquired by over 90% of DCC2 isolates and 3 DCC1 isolates (Figure 16). This suggested this region had possibly recombined with a sublineage of DCC1 and could have contributed to both the emergence of DCC2 and the ongoing spread and adaptation of DCC1 (Figure 14).

Encoded within the first cluster (10071_6#72_00250-10071_6#72_00272) were four *mce* genes, two of which were orthologous to *mce* genes, Rv1970 and Rv1971, in *M. tuberculosis* H37Rv (Figure 16) (230). *Mce* operons usually consist of two *yrbE* genes with homology to ABC transporter permeases, followed by six *mce* genes which share homology to substrate binding proteins (259). Disruption of *mce* operons in *M. tuberculosis* have been shown to cause changes to the virulence of the organism (259). However, given that a complete *mce* operon was not encoded within this cluster of consecutive locus tags, it was unclear what function these *mce* genes were performing, although orphan *mce* genes have been reported (260). The partial *mce* operon, along with several hypothetical proteins and other putatively cell wall associated genes, were flanked by a recombinase (10071_6#72_00251) and a possible transposase (10071_6#72_00266).

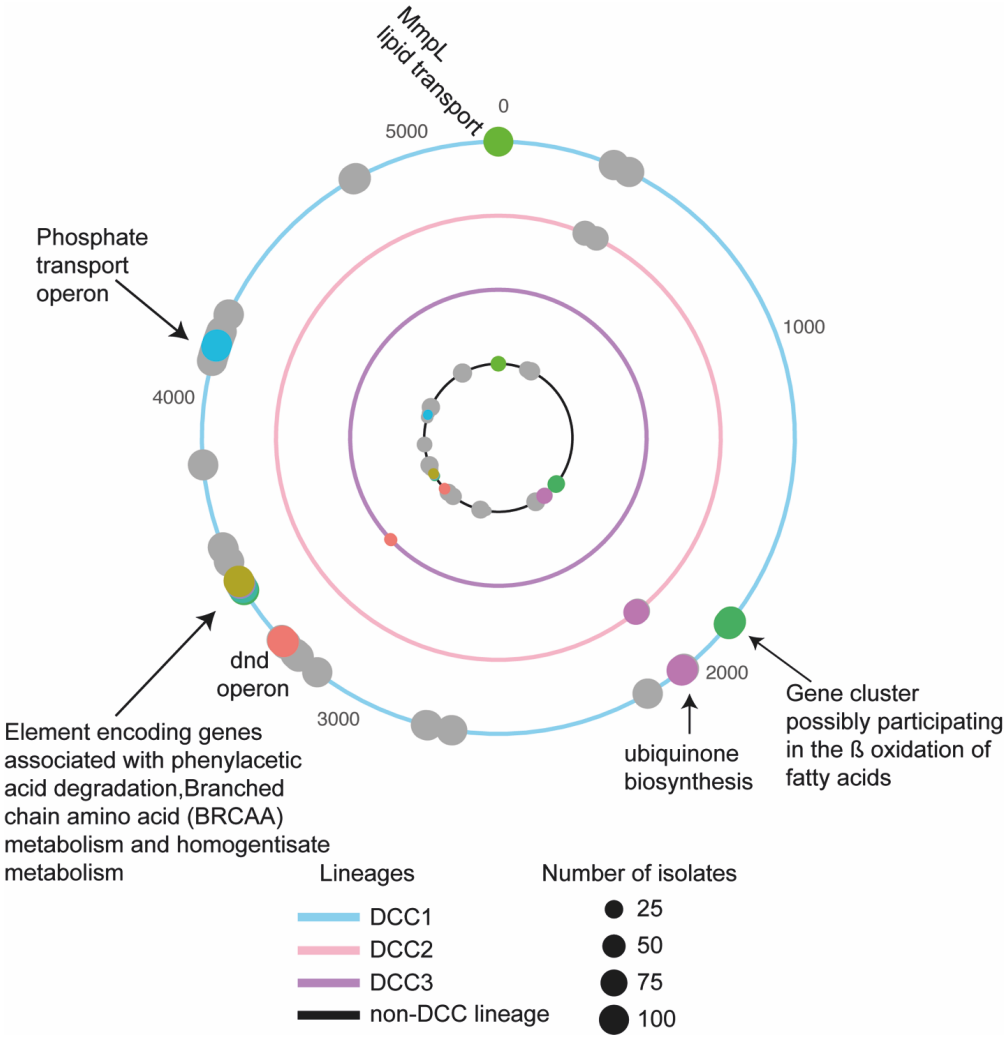


Figure 15: The DCC1 accessory genome includes genes involved in branched chain amino acid transport, biosynthesis and metabolism

Concentric circle plot showing genes present in at least 90% DCC1 isolates (outer circle) and less than 10% of isolates not associated with DCCs (inner most circle). Regions where clusters of genes are associated with DCC1. The functions of many of these genes have been linked with the adaptation of other pathogens to the human host. Colored points are genes or clusters of genes with potentially beneficial functions. Grey points represent genes where deciphering the function of the gene was not possible. The size of the point represents the proportion of isolates encoding the gene.

Amongst the genes in the second cluster (10071_6#72_00289-10071_00316) were genes with similarity to an alkanesulfonates transport system (10071_6#72_00290-10071_6#72_00293) (261), although the gene order in the *E. coli ssuEADCB* operon was not mirrored here (appendix table 1.21). The alkanesulfonate transport system is a mechanism by which organisms cope with cysteine and sulphate starvation by enabling the use of aliphatic sulfonates as a source of sulfur (261).

Multiple AhpC/TSA family proteins and sigma factors were encoded within both clusters. AhpC/TSA subunit C encoding proteins include many antioxidants including the *ahpC* gene of *M. tuberculosis*, which is critical for enabling the survival of *M. tuberculosis* in the oxidative environment of the macrophage (262, 263). Sigma factors also play a role in Mycobacterial pathogens response to stress by causing specific regulatory changes (264). Pfam and InterPro also annotated 10071_6#72_00312 as a stress induced transcriptional regulator. It should be noted, however, that genes with functions potentially associated with energy metabolism (cytochrome c biogenesis) as well as lipid metabolism and cell wall biosynthesis were also encoded within this cluster.

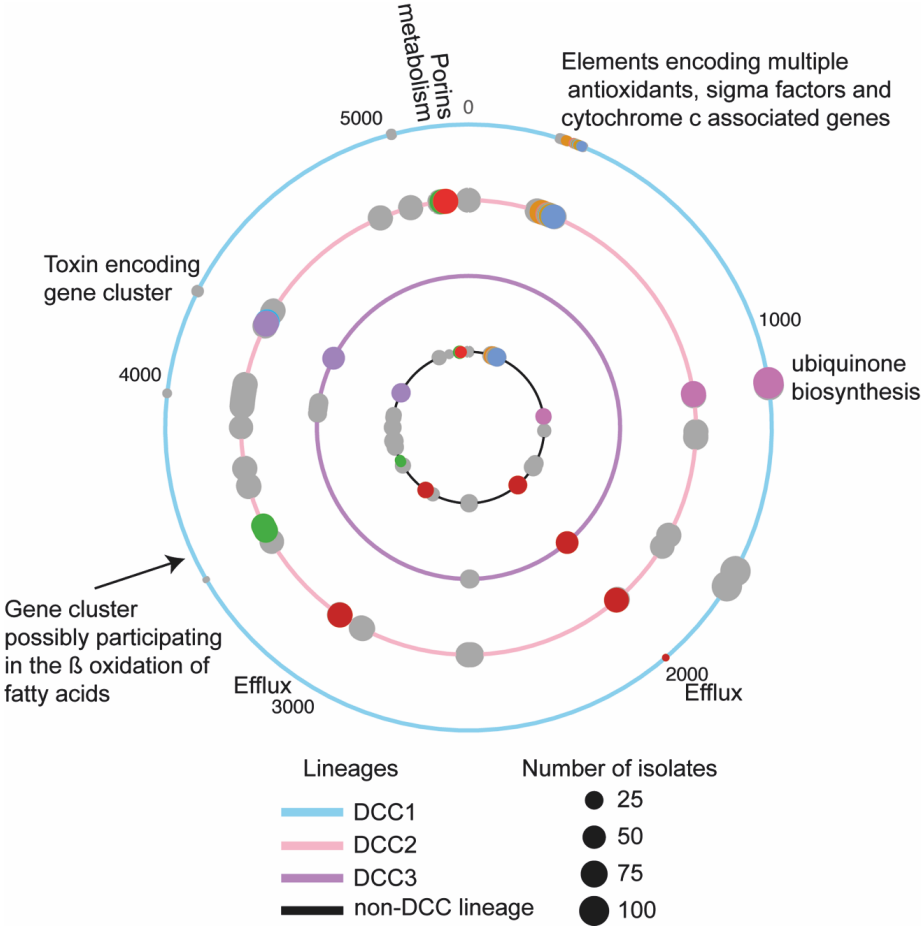


Figure 16: The LCA of DCC2 was enriched with genes associated with metabolism and transport across the membrane

Genes present in at least 90% DCC2 isolates and less than 10% of isolates not associated with DCCs (inner most ring). Overlap between the candidates for DCC2 and the other DCCs was evident. Candidates identified included metabolism associated genes, antioxidants, sigma factors and efflux pumps.

2.3.7.3 DNA degradation operon present in DCC1 and DCC3

A subset (10208_3#20_03268-10208_3#20_03272) of a large cluster of genes (10208_3#20_03269-10208_3#20_03276) present in over 90% of DCC1 isolates were also present in a subset of seven DCC3 isolates (Figure 15, appendix table 1.21 and table 1.23). These genes encode the DNA degradation (*dnd*) locus, which has previously been described in the *M. a. abscessus* ATCC19977 reference genome (82). A *dnd* phenotype could enable these lineages to take advantage of the DNA rich CF sputum as a nutrient source as well as potentially providing resistance to oxidizing agents, as has been observed in *Salmonella enterica* (265).

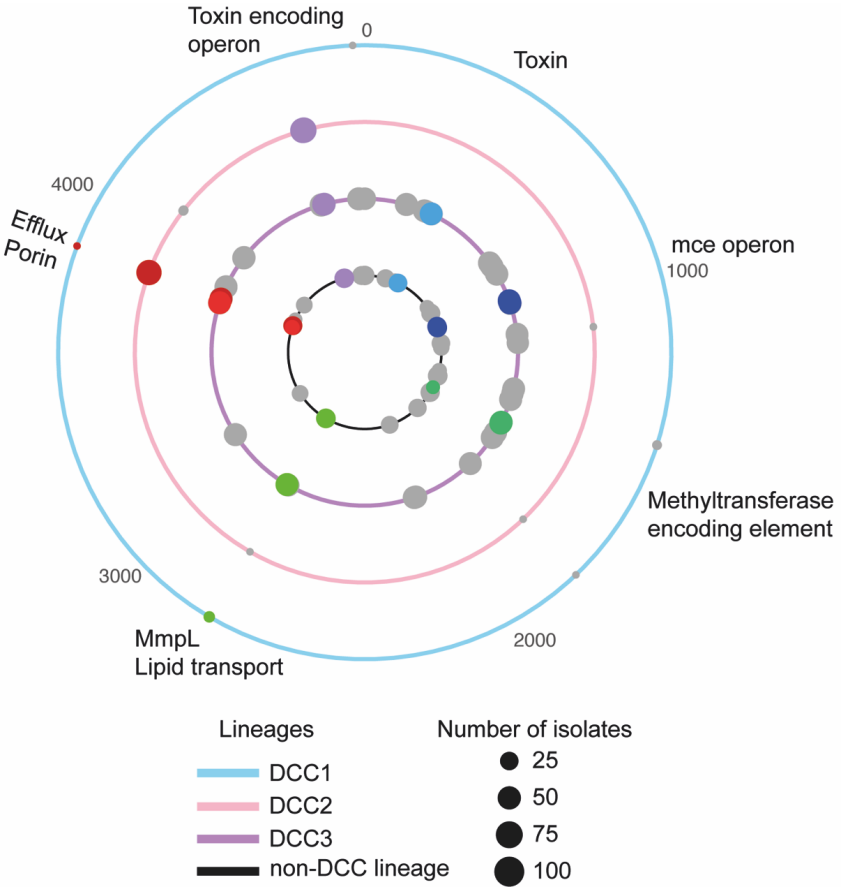


Figure 17: The LCA of DCC3 was enriched with genes associated with the membrane

Genes present in 90% DCC3 isolates and less than 10% of isolates not associated with the DCC lineages. Amongst the genes acquired by DCC3 was a gene involved in efflux, the same toxin encoding element as was present in DCC2, a complete *mce* operon and a mobile element encoding a methyltransferase.

2.3.7.4 Admixing of phage associated genes between DCC2 and DCC3

Three of the clusters of genes identified as present in 90% of DCC2 isolates had gene content indicative of phage, although only one (10071_6#72_03926-10071_6#72_03988) produced a hit against the PHASTER database. The 82 genes encoded by these three gene clusters included 59 (72%) genes for which the annotation could not be determined. However, within each of these three clusters a subset of genes were also found to be present in a proportion of DCC3 isolates (appendix table 1.22).

Within one of the clusters (10071_6#72_04228-10071_6#72_04239), the subset of genes (10071_6#72_04236-10071_6#72_04239) that were also present in a large proportion of DCC3 isolates included a gene (10071_6#72_04238) which was predicted by Pfam and InterPro to encode a toxin domain of the sort found in polymorphic toxin systems (Figure 16, 17) (266). The machinery characteristically encoded with the toxin encoding genes in order for the toxin to be exported was not present within the gene cluster or up or down stream of the cluster and thus it was more likely to form part of a toxin/antitoxin system (266).

2.3.8 Functional similarity between the gene clusters unique to each DCC

2.3.8.1 Genes with metabolism and transport functions prominent in DCC1 gene clusters

DCC1 encoded 178 genes not present in any other DCC isolates and only present in less than 10% of isolates that were not part of the DCC lineages (appendix table 1.21). Amongst these were several clusters of genes which could have provided an advantage to the lineage in establishing itself as a prevalent MABSC lineage in the CF community.

DCC1 encoded a cluster of genes potentially involved beta oxidation of fatty acids (10208_3#20_01853-10208_3#20_01862). One of the genes within this cluster, 10208_3#20_01855, was orthologous to *M. tuberculosis* H37Rv gene Rv2790c (*lpt1*), which was annotated as a lipid-transfer protein (230). As previously mentioned, beta oxidation of fatty acids is a key mechanism by which *M. tuberculosi*s remodels its metabolism to cope with the different carbon sources available in the host.

Interestingly, one of the largest clusters of genes (10208_3#20_03451-10208_3#20_03481) included genes that were identified through SNP density analysis to have acquired a significantly different number of nonsynonymous SNPs on the branch leading to the LCA,

such as the BCKAHD enzyme complex (244). Also amongst this cluster was the homogentisate metabolism operon and the phenylacetic degradation pathway (Figure 15). These systems, which are not typically encoded by mycobacteria, have been previously described within the *M. a. abscessus* ATCC19977 reference genome (82). Homogentisate catabolism has been linked with the adaptation of *P. aeruginosa* to the CF lung, whilst phenylacetic acid degradation has been associated with the adaptation of *Burkholderia cepacia* to the lung (179, 267).

A further large cluster of genes (10208_3#20_04114 - 10208_3#20_04195), included a complete transposon, encoding genes with energy metabolism associated functions, followed by genes associated with arsenic transport and the genes that encode the phosphate transport operon, *pstSCAB* (Figure 15). Two of the four genes which make up this transporter, *pstS* and *pstC* were orthologous to the corresponding genes in the *M. tuberculosis* H37Rv *pstSCAB* operon, Rv0928 and Rv0929. This operon plays a key role in phosphate homeostasis and as phosphate is an essential nutrient for cellular functions, it is key to survival within the host (268, 269).

2.3.8.2 DCC2 accessory genome encoded fatty acid metabolism and transport related genes

120 genes were present in only DCC2 isolates and less than 10% of isolates not part of DCC lineages. Similarly to DCC1, this included a cluster (10071_6#72_03465-10071_6#72_03485) of genes with functions that would fit with their participation in the beta oxidation of fatty acids (Figure 16). A cluster of genes (10071_7#72_05035-10071_7#72-05045) potentially involved in energy and respiration metabolism were also encoded.

Several genes with functions associated with transport across the cell membrane were present within the candidates associated with the emergence of DCC2. These included two (10071_6#72_05050-10071_6#62_05051) mycobacterium smegmatis porin (*mspA*) genes. Porins have been associated with the resistance displayed by *M. chelonae* to aldehyde based disinfectants (270). A MFS protein (10071_6#72_03057) was also present on the candidate list. MFS proteins have been linked with drug efflux and transport of essential nutrients and metal ions (271).

2.3.8.3 DCC3 acquired a complete mce operon and a mobile element encoding a methyltransferase

101 genes were encoded by over 90% of DCC3 isolates and less than 10% of isolates not associated with a DCC and no other DCC isolates. Amongst these genes were a complete *mce* operon. *Mce* operons gained their name due to the loss of the first gene in the sequence (*mceA*) of six preventing transformed *E.coli* from being taken up by macrophages and HELA cells (259, 272). *M. tuberculosis* encodes four *mce* operons which have been shown to be differentially expressed in different growth stages and have been linked with its pathogenicity (272, 273). Interestingly, none of the *mce* related genes within this cluster were found to be orthologous to genes in *M. tuberculosis* H37Rv (Appendix table 1.23).

A cluster of genes (10208_3#26-01553-10208_3#26-01558) which consisted of an integrase followed by a recombinase, a resolvase, an N6 adenine specific S-adenosyl-L-methionine dependent methyltransferase (*dpnM*), a DGQHR domain encoding gene and a ribonucleotide reductase (*rnr*), was also identified amongst the candidates associated with the emergence of DCC3. Methyltransferases, by modifying a specific DNA base within a conserved motif, can cause changes in regulation which can have consequences on the pathogenic potential of organisms (214, 274). Ribonucleotide reductases perform the essential step converting the four ribonucleotide triphosphates into their corresponding deoxyribonucleotide triphosphates (275). It was also interesting to note that similarly to DCC2, an *mspA* porin gene was present on the candidate list for DCC3, although only one as opposed to two.

To see if the evidence of functional overlap, for example the presence of multiple genes associated with the beta oxidation of fatty acids in DCC1 and DCC2 or the porins encoded by DCC2 and DCC3, was statistically significant, GO-term enrichment and pathway analysis was carried out. A combination of TopGO and Blast2GO, was used to determine whether the DCCs were enriched with the same functions in comparison to other MABSC lineages or whether changes had occurred in the same or connected pathways.

2.3.9 Accessory genomes of the DCCs not statistically significantly enriched with the same GO-terms

InterProScan was used to annotate the 18,386 genes which were present in 1-95% of the 512 MABSC isolates used in the pangenome analysis with GO-terms. However, only 4,137 (22%) of these genes were assigned a GO-term and only 558 (202 BP GO-terms, 20 CC GO-terms, 336 MF GO-terms) unique GO-terms were identified (Appendix table 1.24). Furthermore, only 77 of the 183 candidates identified for DCC1, 24 of the 217 candidates

identified for DCC2 and 35 of the 119 DCC3 candidates were assigned at least one GO-term.

Despite the fact that not all the genes were assigned a GO-term, which meant that if any functional enrichment was identified it would not represent the complete functional picture of the DCCs, a test to see if any of the DCCs were statistically significantly enriched with particular GO-terms was carried out using the Fisher's exact test, applied via TopGO. However, the DCCs were not found to be statistically significantly enriched (p-value 0.01) with any GO-terms from either the MF, BP or CC ontologies (Appendix tables 1.25-1.33).

Blast2GO (v.4.1.9) was used to map the candidate genes for each DCC to the KEGG pathway database. Of the 183 genes which characterized the early ancestral lineages of DCC1, 10 (5%) were assigned to a total of 24 KEGG pathways. 20 (5%) of the 217 genes associated with the emergence of DCC2 were mapped to 21 KEGG pathways and four (3%) of the 119 genes associated with the emergence of DCC3 were assigned to seven KEGG pathways (Figure 18). In total, the 34 candidate genes where an enzyme code was able to be assigned were mapped to 39 pathways (appendix table 1.34).

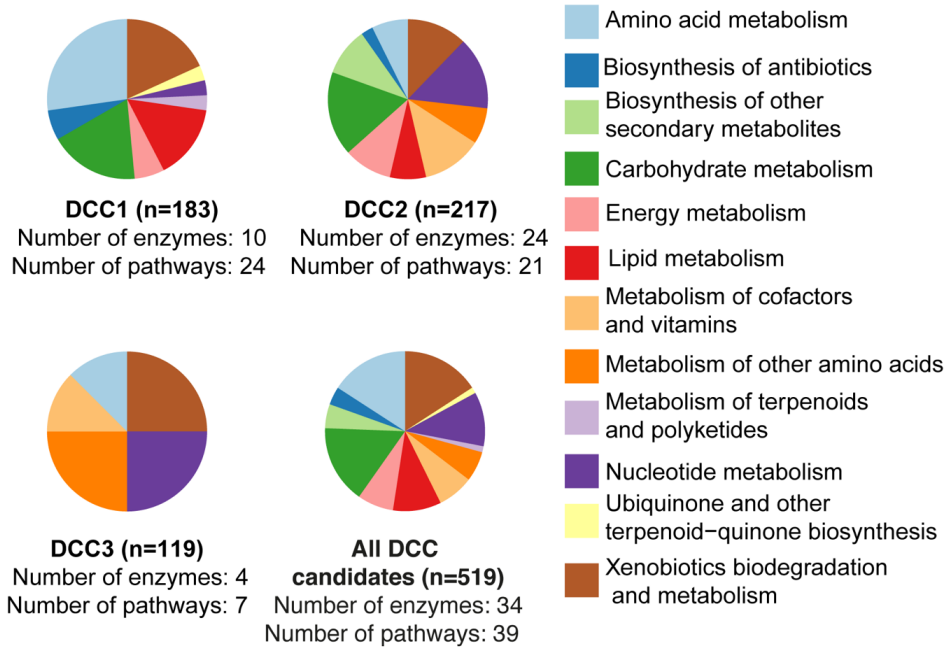


Figure 18: Incomplete functional picture gained from KEGG pathway analysis

Blast2GO assigned 5% of the candidate identified for DCC1 and DCC2 and 3% of the candidates for DCC3 to KEGG Pathways. The 10 DCC1 candidates were assigned to 9 functional areas, with amino acid metabolism pathways the most common. The 24 DCC2 candidates were assigned to 10 functional areas, with nucleotide and carbohydrate metabolism the most common. The 3 DCC3 candidates were assigned to 5 functional areas, with xenobiotic biodegradation and metabolism, nucleotide metabolism and metabolism of other amino acids being the most common. Each DCC acquired candidates potentially functioning in amino acid metabolism pathways and xenobiotic biodegradation and metabolism pathways.

Pathways involved in amino acid metabolism and xenobiotics biodegradation and metabolism were the most common, with eight pathways within each of these categories having at least one candidate gene mapping to them (Figure 18). Lipid (7) and Carbohydrate (5) metabolism pathways also accounted for a large proportion of the pathways detected. Candidate genes from all the DCCs mapped to pathways involved in amino acid metabolism, nucleotide metabolism, lipid metabolism, Xenobiotics biodegradation and metabolism, and metabolism of cofactors and vitamins. Only candidates from DCC1 and DCC2 mapped to carbohydrate metabolism and energy metabolism pathways. A single candidate for DCC1 mapped to a pathway involved in the metabolism of terpenoids and polyketides and a single candidate from DCC2 mapped to a pathway associated with biosynthesis of other secondary metabolites. Whilst this could suggest that the early ancestral lineages of each of the DCCs had undergone gene content changes in pathways involved in similar functions, the low number of candidates mapped to a pathways and the number of candidates which were

mapped to multiple pathways made it impossible to confidently draw any conclusions from KEGG pathway analysis (Figure 18, Appendix table 1.34).

Through the pangenome analysis it was evident that each of the DCCs had acquired potential virulence factors either just before or just after their LCA, some of which were also present in other DCC isolates as well as other lineages which had either not increased in prevalence within the CF community or not to such an extent. This suggested that it was the unique combination of virulence factors acquired by each of the DCCs that potentially gave these lineages an advantage over other MABSC lineages in the CF lung environment, with advantages in metabolic flexibility and transport across the membrane functional areas in which the DCCs had potentially been better adapted to survive in the human host.

To begin to further investigate the candidates identified through these analyses, the methyltransferase encoding mobile element present in the DCC3 lineage (see section 2.3.7.3) was selected for further analysis due to its presence in the DCC responsible for the Papworth and Seattle CF center outbreaks as well as the epidemic of SSTIs in Brazil, and the fact that it was present in only three isolates that were not part of DCCs.

2.3.10 Preliminary functional validation of the methyltransferase encoding mobile element

The sequence of the mobile element encoded by a representative of the DCC3 lineage, BIR1049, was selected as a reference for the raw reads of all the global population and publicly available MABSC isolates used in this study to be mapped to. This confirmed the presence of the mobile element in all 57 DCC3 isolates and just three further MABSC global population isolates, two *M. a. massiliense* isolates, AUS791 and DEN538, and one *M. a. bolletii* isolate, RHS37 (Figure 19).

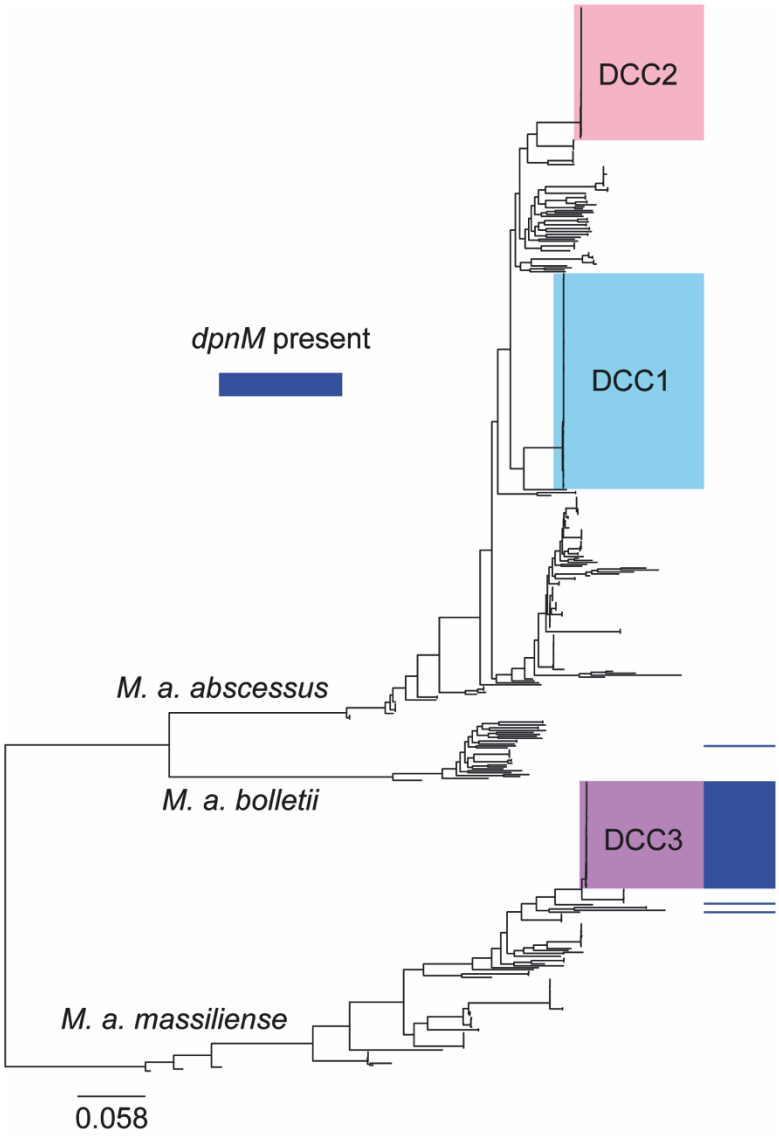


Figure 19: dpnM is encoded by all DCC3 isolates and three further MABSC isolates

The MABSC global population structure with the metadata column representing the isolates which encoded a methyltransferase, dpnM. All DCC3 isolates were found to encode *dpnM*, along with two further *M. a. massiliense* isolates and an *M. a. bolletii,* none of which belonged to recently expanded lineages. The presence of the dpnM was determined by mapping all the MABSC isolates to a representative sequence of the MABSC, encoded by DCC3 isolate BIR1049.

Nucleotide blast comparisons between the sequence of the mobile element encoded by BIR1049 and the outliers showed that the mobile element was inserted in the same position in all the isolates and bookended by two 42 base pair direct repeats, suggesting that the mobile element had been introduced into the genomes via site specific recombination (Figure 20). This also showed that the *rnr* gene present in the gene cluster was not part of the element as it was encoded outside the direct repeats. Thus, the mobile element encoded an

integrase, a resolvase[4], dpnM, and a hypothetical protein with a DGQHR domain (Table 6). A comparison of this hypothetical protein against the NCBI RefSeq database identified regions of homology with the *dndB* family domain as well as a possible YraN endonuclease domain.

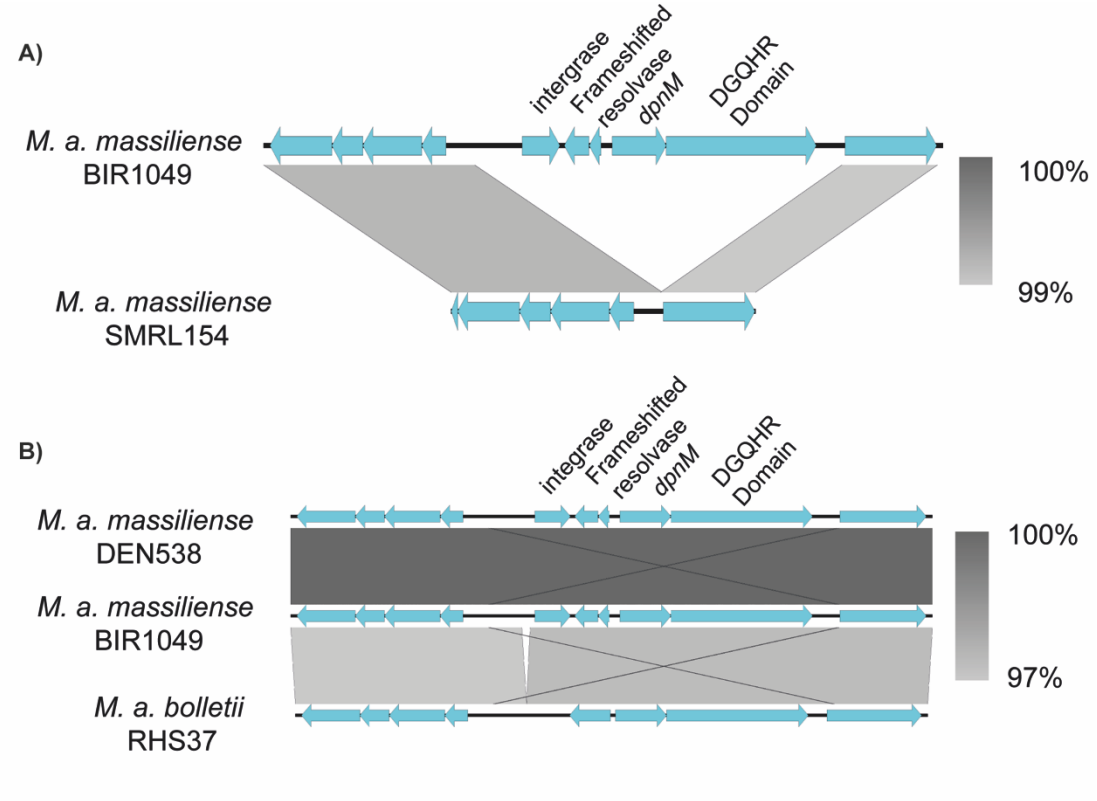


Figure 20: dpnM encoding mobile element inserted between direct repeats

Blastn comparisons, produced using EasyFig(276), which A) shows the insertion of the mobile element in comparison to the most closely related isolate where was not present and B) shows the high sequence identity between the inserted element found in *M. a. massiliense* (DEN538) and *M. a. bolletii* (RHS37) isolates and the representative DCC3 isolate (BIR1049).

Comparing the protein sequence of dpnM against the REBASE database indicated that it was likely to be recognizing a GATC motif and modifying the amino group at the C-6 position on the adenine (Table 6). This comparison also highlighted two conserved motifs within the active site of dpnM: the FXGXG motif within the AdoMet binding site and the DPPY motif, present within the catalytic domain (277). Mutations affecting the aspartate and tyrosine sites of the DPPY motif and the phenylalanine site of the FXGXG motif have been shown to result in the loss of function of the methyltransferase (237, 238).

[4] The resolvase gene encoded by the mobile element in BIR1049 (121632_2#22_011822-011823), DEN538, AUS791 was represented by two genes in the annotations of these isolates due to a frameshift caused by the deletion (GAGgGTG) of a guanine base in codon 60 of CDS 12163_2#22_01182 (BIR1049 example)

In order to determine the structure of dpnM and investigate its catalytic nature, structural modelling of dpnM was carried out by Dr Sony Malhorta.

2.3.10.1 Structural analysis of dpnM

Using PDB ID-2dpm chain A of *Streptococcus pneumoniae* as a template (Figure 21Ai), Dr Sony Malhotra confirmed the presence of the two conserved motifs, FXGXG and DPPY, in the ado-met binding site and catalytic domain respectively (Figure 21Aii). Mutations in these motifs, F42S in FXGXG, D184A and Y187L in DPPY, have previously been shown to have inactivating effects (237, 238). These mutations were introduced into the model to investigate their impact on the function of dpnM.

Dr Sony Malhorta predicted that F42S would have an inactivating effect due S42 disrupting the π-π interactions between F42, the adenine ring of SAM and F168 (Figure 21B.i and Figure 21B.ii). D184A was also predicted to have an inactivating effect. This was because the D184A mutation was predicted to disrupt the electrostatic interactions formed between D184, the nitrogen of the amino group and K19 (Figure 21C.i and Figure 21C.ii). Contrastingly, Y187L was predicted to have a destabilizing as opposed to inactivating effect. This was because the H-bond interactions with A265 were retained when Y187 was replaced with L187 but the stacking with the aromatic ring of the flipped target adenine base was lost.

Table 6: Annotations of the CDSs encoded by the mobile element present in all DCC3 lineage isolates

Locus tag	Prokka	Pfam	Interpro	NCBI	REBASE
12163_2#22_01181	Integrase fusion protein	no-hits	No hits	refseq: C-terminal catalytic domain of integrase from bacterial phages and conjugate transposons	NA
12163_2#22_01182[3]	Site-specific recombinase, DNA-invertase hin, N-terminal domain	Resolvase N-terminal domain and HTH DNA binding domain	Resolvase N-terminal domain and HTH DNA binding domain	Swissprot: invertase/resolvase subfamily Serine recombinase family	NA
12163_2#22_01183[3]	Putative resolvase, N-terminal domain	NA	NA	NA	NA
12163_2#22_01184 (dpnM)	dpnIIA, DNA adenine methylase	D12 N6 adenine-specific DNA methyltransferase	D12 N6 adenine-specific DNA methyltransferase	Swissprot: DNA-adenine specific methyltransferase (N6-adenine specific)	Recognition site: GATC Conserved domains: FXGXG and DPPY
12163_2#22_01185	DGQHR domain containing protein	DNA sulphur modification associated - conserved DGQHR domain	DGQHR containing domain, similar to DNA sulphur modification protein DndB	refseq: DGQHR domain, uncharacterized conserved domain. Some proteins have been found to be part of DNA phosphorothioation systems, YraN domain: predicted endonuclease	NA

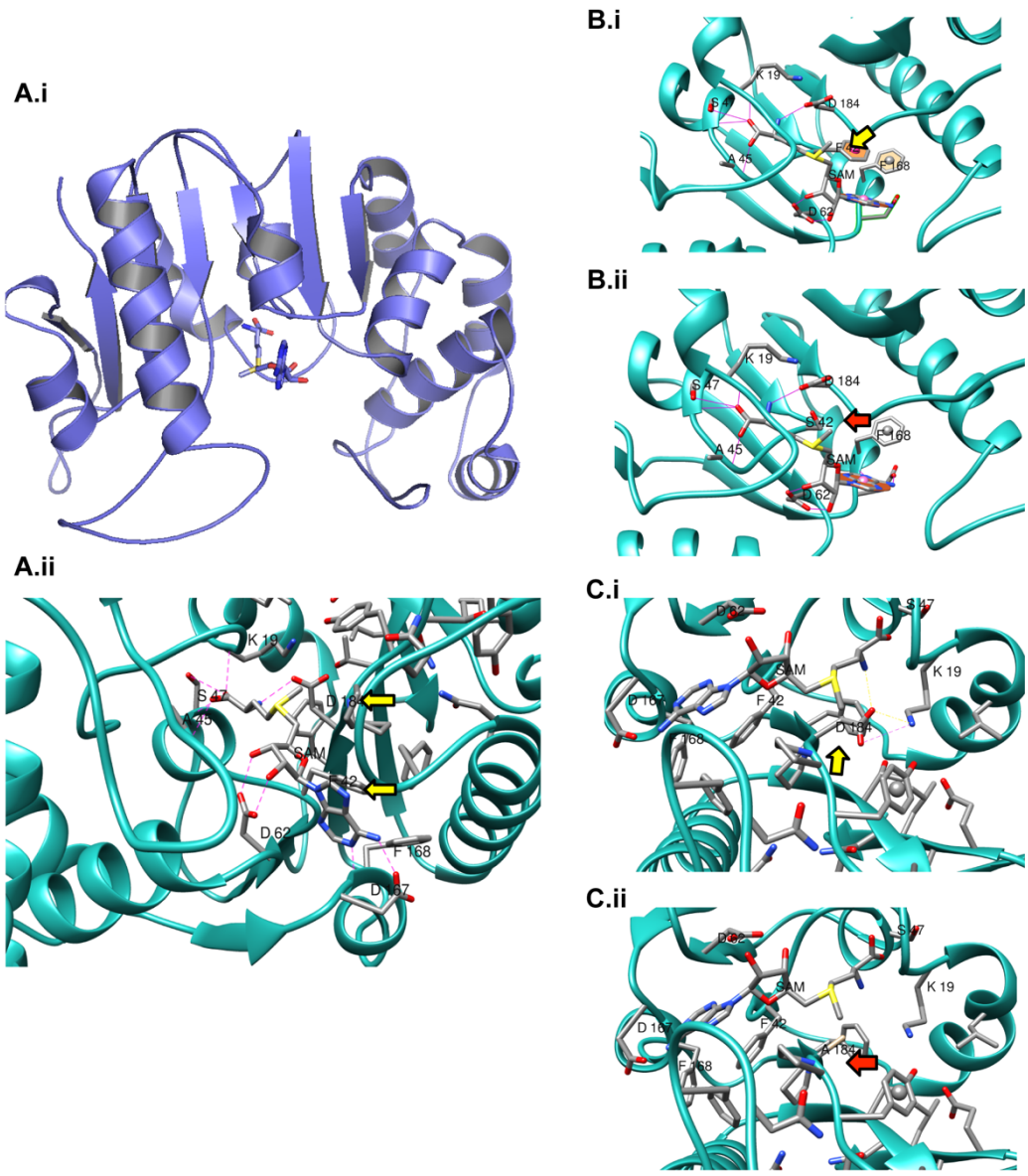


Figure 21: Structural modelling of dpnM

A.i) Structural model of the methyltransferase determined using PDB ID-2dpm chain A of *Streptococcus pneumoniae.* A.ii) The binding pocket of the methyltransferase when S-adenosyl-L-methionine (SAM) is bound. The conserved amino acids F42 of the FXGXG and D184 of DPPY motifs are indicated by the yellow arrows. B.i) Structural modelling of the interaction between F42 and SAM (yellow arrow) and B.ii) the effect of the inactivating mutation F42S (red arrow). C.i) Structural modelling of the interaction between D184 and SAM (yellow) with C.ii) showing the effect of the D184A inactivating mutation. Figures curtesy of Dr Sony Malhotra.

2.3.10.2 dpnM recognizes and modifies an RGATC motif:

Next, in order to confirm that the motif modified by the dpnM was GATC and to examine the methylome of the DCC3 lineage, SMRT sequencing was undertaken on the isolates recorded in Table 1[5]. Modification and motif analysis showed that all the wild type (WT) isolates which encoded dpnM were predicted to encode a modified adenine bases within a GATC motif (Table 7). Although, the specificity of the bases flanking the motif differed between DEN538 (RGATCC) and RHS37 and BIR1049 (both BRGATCC). No other motifs were modified by all the isolates encoding dpnM. The modification of this motif by dpnM was further supported by the absence of this modified motif in the BIR1049 knock out (KO), as well as its absence in the WT isolates which did not encode dpnM, although this could not be conrfirmed for SMRL154 as it was not successfully sequenced (Table 7). The motif could also not be confirmed through complementation, with a modified RGATCC motif not detected after SMRT sequencing of the BIR1049Δinsertion:dpnM strain (Table 7).

The proportion of the motifs encoded by each of the genomes that were modified differed. 47% of the 4,879 BRGATCC motifs encoded by BIR1049 were modified, 73% of the 5,653 RGATCC encoded by DEN538 were modified and 62% of the 5,005 BRGATCC encoded by RHS37 were modified (Table 7), suggesting the BRGATCC motif was likely to be an overprediction by the software, and that RGATCC is the true recognition motif in all of the genomes.

In order to identify whether the modification of the RGATCC motif causes changes in gene expression that could be linked to virulence and thus clarify whether the insertion of the dpnM played a role in the preadaptation of DCC3, the genomic positions of the modified motifs need to be identified and correlated with the results of differential expression analysis after RNA sequencing of the BIR1049 WT, BIR1049Δinsertion and BIR1049Δinsertion:dpnM. This work is in progress.

It was also interesting to note that as well as the modified GATC motif, two pairs of non-palindromic, bipartite asymmetric modified motifs were detected in the BIR1049 WT and BIR1049 knockout strains, which suggested other methyltransferases were encoded by BIR1049 (Table 7). This suggested that the methylome of BIR1049 and potentially that of the DCC3 lineage was unique when compared to that of the four other lineages tested. Further modified motifs were also detected in the genetically distinct isolates that encoded dpnM,

[5] SMRL154 failed SMRT sequencing

with an asymmetric, bipartite modified motif identified in DEN538 and two non-palindromic modified motifs identified in RHS37. Contrastingly, in the isolates selected for SMRT sequencing which didn't encode dpnM, no modified bases were detected, which suggested that no methyltransferases were present in these genomes.

2.3.10.3 dpnM is potentially playing a role in intracellular survival

The three DCCs and other densely clustered although less expanded MABSC lineages were found to be more successful at infecting human-derived macrophages and surviving within these cells in comparison to MABSC lineages from genetically distinct backgrounds (73). To investigate whether dpnM was contributing to these phenotypes, Daniela Rodriguez-Rincon performed the original assays described by Bryant et al (2016) on the following isolates: BIR1049wt, BIR1049Δinsertion, BIR1049Δinsertion:dpnM, BIR1049Δinsertion:dpnM(D184A), BIR1049Δinsertion:dpnM(Y187L) and BIR1049Δinsertion:dpnM(F42S).

No significant differences in phagocytic uptake were observed by Daniela Rodriguez-Rincon between the BIR1049 strains (Figure 22A), however, BIR1049Δinsertion was found to be unable to survive within monocyte derived human macrophages, which suggested that dpnM could be playing a role in intracellular survival (Figure 22B). The complemented strains BIR1049Δinsertion:dpnM and BIR1049Δinsertion:dpnM(Y187L) restored intracellular survival whereas the strains complemented with versions of dpnM predicted to have inactivating mutations, BIR1049Δinsertion:dpnM(D184A) and BIR1049Δinsertion:dpnM(F42S), failed to restore intracellular survival. As well as providing further evidence that dpnM could be playing a role in intracellular survival, these results also confirm Dr Sony Malhotra's predictions of the impact of these mutation based on the structural modelling of dpnM (Figure 21). Whilst these results suggest that dpnM is playing a role in intracellular survival, these results are preliminary as thus far these findings have been unable to be confirmed.

Table 7: Summary of the modified motifs detected in the seven isolates that underwent SMRT sequencing

Isolate	Mobile element dpnM	Predicted motif	Partner motif	Modification type	% motifs modified in genome	# of motifs detected	# of motifs in Genome	Mean modification QV	Mean Motif Coverage
M. a. massiliense BIR1049	present	GAGNNNNNNGTTG	CAACNNNNNNCTC	m6a	70.90	441	622	44.56	28.62
		CAACNNNNNNCTC	GAGNNNNNNGTTG	m6a	66.08	411	622	44.19	28.43
		GAACNNNNNNTCC	GGANNNNNNGTTC	m6a	68.93	579	840	44.21	28.71
		GGANNNNNNGTTC	GAACNNNNNNTCC	m6a	53.10	446	840	42.51	29.68
		BRGATCC	NA	m6a	46.94	2290	4879	44.03	29.61
M. a. massiliense BIR1049 Knockout	absent	GAACNNNNNNTCC	GGANNNNNNGTTC	m6A	98.21	824	839	72.56	47.03
		GGANNNNNNGTTC	GAACNNNNNNTCC	m6A	91.06	764	839	66.64	47.76
		CAACNNNNNNCTC	GAGNNNNNNGTTG	m6A	97.44	609	625	73.37	46.57
		GAGNNNNNNGTTG	CAACNNNNNNCTC	m6A	95.84	599	625	71.55	47.03
BIR1049 Complemented Wild type	present	GAACNNNNNNTCC	GGANNNNNNGTTC	m6A	90.70	761	839	57.86	33.07
		GGANNNNNNGTTC	GAACNNNNNNTCC	m6A	83.67	702	839	55.47	34.07
		GAGNNNNNNGTTG	CAACNNNNNNCTC	m6A	89.08	555	623	58.53	33.35
		CAACNNNNNNCTC	GAGNNNNNNGTTG	m6A	88.28	550	623	57.8	33.09
M. a. massiliense SMRL154	absent	NA	NA	NA	NA	NA	NA	NA	NA
M. a. massiliense DEN538	present	GGANNNNNTCC	GGANNNNNTCC	m6A	98.50	2563	2602	117.08	79.5
		RGATCC	NA	m6A	72.58	4103	5653	88.2	79.52
M. a. massiliense AUS856	absent	NA	NA	NA	NA	NA	NA	NA	NA
M. a. bolletii RHS37	present	GARCCAG	NA	m6A	98.83	1691	1711	117.36	74.66
		GCCCGAG	NA	m6A	96.24	1638	1702	106.59	75.55
		BRGATCC	NA	m6A	62.44	3125	5005	72.4	75.43
M. a. bolletii DEN515	absent	NA	NA	NA	NA	NA	NA	NA	NA

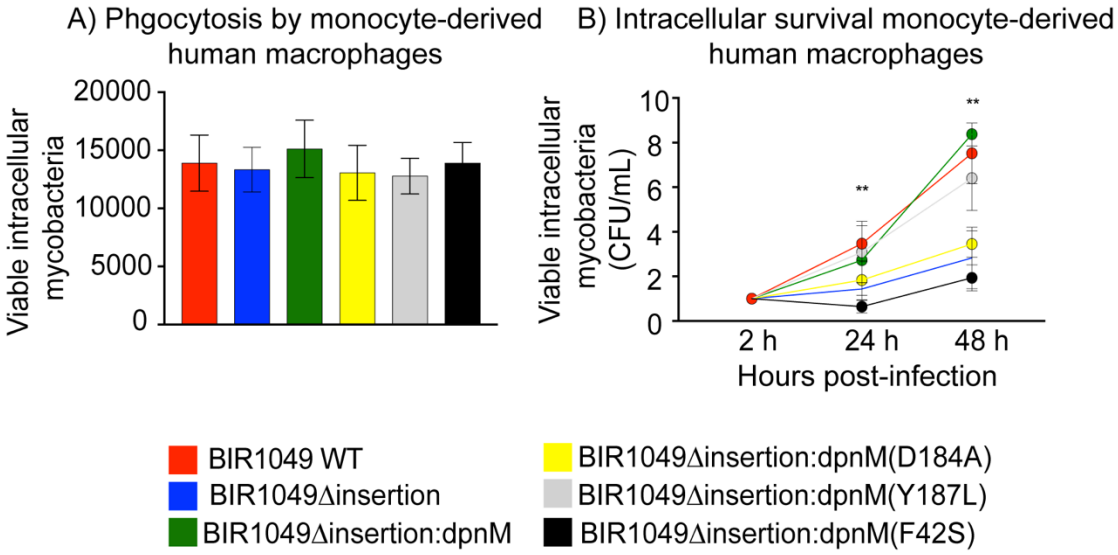


Figure 22: dpnM is potentially playing a role in intracellular survival

A) No significant difference in phagocytic uptake was observed between the BIR1049 strains with a functioning dpnM (*M. a. massiliense* BIR1049wt, BIR1049Δinsertion:dpnM, BIR1049Δinsertion:dpnM(Y187L)) and without a functioning dpnM (BIR1049Δinsertion), BIR1049Δinsertion:dpnM(D184A) BIR1049Δinsertion:dpnM(F42S)). **B)** Significant differences (*** p <0.001) in intracellular survival were observed between the three strains, BIR1049wt, BIR1049Δinsertion:dpnM and BIR1049Δinsertions:dpnM(Y187L), that encoded a functioning or destabilized dpnM compared to the three strains where dpnM was not present. Figures adapted from those provided by Daniela Rodriguez-Rincon

2.4 Discussion

The rapid emergence, global spread and greater virulence of clustered MABSC lineages led to the hypothesis that these lineages could have been preadapted to the CF lung environment (73). The research in this chapter aimed to investigate this hypothesis by examining the genetic changes that occurred on the branches leading to the LCA of the three largest clustered lineages, the DCCs (Figure 7). The genetic changes that occurred on these branches could indicate either changes that pre-adapted these lineages to the CF lung environment or changes that were rapidly fixed in the lineages after their initial invasion of the CF lung niche, but before their clonal expansion. By identifying these changes it may be possible to gain a greater understanding of what has driven the emergence of the more prevalent MABSC lineages as well as increase our understanding of MABSC virulence factors. The results of this chapter showed that no single genetic factor was responsible for the emergence of the DCCs, which suggests that the DCCs have expanded in the CF

community due to increased opportunity. However, the functions of the genes identified as enriched in the LCA of the DCCs did provide evidence as to why it was these lineages that have expanded and highlighted functional areas that could contribute to the virulence of the MABSC.

A clear indication that prior to their clonal expansion the DCCs had undergone rapid adaption to a new environment would have been if the SNPs on the branches leading to the LCA of the DCCs were under positive selection. However, this was not observed, with the branches leading to the LCA of the DCCs all under strong purifying selection (Figure 8). In fact the dN/dS values on the branches leading to the LCA of the DCCs were very similar to those observed between other mycobacterial species as well as other species within the actinobacteria phylum (278). This suggested that prior to their LCA, the DCCs were evolving under the selective constraints applied by an environment to which they were already adapted. Contrastingly, over shorter time scales, including many of the branches representing the evolution of the DCC lineages after their LCAs, the dN/dS values increased (Figure 8). Whilst this could indicate that strong positive selection was occurring after the LCAs of the DCCs, it could also represent the fact that comparisons between closely related isolates can result in high dN/dS values because purifying selection has not had time to occur (279).

The results of the dN/dS analysis suggested that if adaptation had occurred prior to the clonal expansion of the DCCs, it was through other evolutionary processes such as more subtle changes in the selection pressure acting upon core genes or changes in gene content. Through both SNP density and pangenome analysis genes potentially associated with emergence of the DCCs were identified. Each of the candidate lists generated consisted of many candidate genes encoded by consecutive locus tags (Appendix table 1.1, 1.2, 1.3, 1.21, 1.22, 1.23). In the case of the SNP density analysis, as shown in Figure 9A and C, there were two peaks representing the consecutive locus tags amongst the candidates list for DCC1 and DCC3. Given that recombination was not removed from the phylogeny these clustered SNPs could be indicative of a recombination event. The high number of synonymous SNPs also acquired by the genes that were found to have gained a significantly different number of nonsynonymous SNPs (Appendix table 1.1) also suggests that these regions could have been acquired via recombination, as nonsynonymous SNPs in recombinant regions often have had time to be purified in the donor organism (280). Similarly, the pangenome candidate lists consisted of multiple clusters of genes with consecutive locus tags, suggesting they had been acquired together (Figure 15, 16, 17). The

genes that bookended some of these clusters, such as recombinases and integrases, supported the idea that these were mobile elements (appendix table 1.21, 1.22, 1.23). The fact that some of the candidates could have been acquired *via* recombination does not rule out their potential role in predisposing the DCCs to be successful in the human host and the functions of some of the candidates overlapped with functional areas associated with adaptation to the human host in other pathogens.

No candidates were identified through SNP density or pangenome analysis that were associated with the emergence of all three DCCs which suggested that the acquisition of a single gene or mobile element or a significantly different SNP density in the same gene had not contributed to the emergence of all three lineages. Whilst there are examples of the emergence of distinct epidemic lineages being driven by the acquisition of the same gene content changes and SNPs, such as the emergence of two distinct epidemic lineages of *Clostridium difficile* (281), there are also examples of epidemic lineages emerging from distinct genetic backgrounds with no shared genetic cause to explain their emergence, such as the most common disease-associated *Legionella pneumophila* strain types (187).

There was evidence, however, of overlapping accessory gene content between two of the three DCCs, which suggests that the acquisition of the same genetic material had potentially contributed to the emergence of the DCCs. Both DCC1 and DCC2 acquired an operon encoding genes that catalyze a step in ubiquinone biosynthesis (Figure 15, Figure 16). UbiD and UbiX catalyze the breakdown of 3-octaprenyl-4-hydroxybenzoate (OHB) to 3-octaprenylphenol (OPP) (258). Ubiquinone is a plasma membrane associated molecule which acts as an electron carrier between the electron donor, which can be NADH dehydrogenase, succinate dehydrogenase or lactate dehydrogenase, and the acceptor, which can be cytochrome oxidases or reductases (282). Thus both DCC1 and DCC2 acquired a gene cluster associated with respiratory metabolism. Ubiquinone has also been hypothesized to play a role in oxidative stress, potentially protecting organisms including *M. tuberculosis*, *P. aeruginosa* and *Salmonella typhimurium*, from the increased levels of reactive oxygen species (ROS) induced through the degradation of long chain fatty acids (283). Therefore, it is possible the presence of this gene cluster could potentially be contributing to the ability of DCC1 and DCC2 to survive longer within macrophages.

Over 90% of DCC2 and DCC3 isolates as well as one DCC1 isolate encoded a possible drug efflux pump (Figure 16, 17). The significantly densely clustered MABSC lineages were found to be more antibiotic resistant than unclustered lineages, however, the genetic determinants

for these, SNPs in the 23s rRNA and 16s rRNA genes, have been identified (73). However, multidrug resistance efflux pumps are known to perform roles not associated with resistance, such as exporting biocides and toxic metals which could be how this efflux pump could be contributing to the success of the DCC lineages (271).

Other examples of overlapping gene content between the DCCs were observed, although these examples consisted of genes present in over 90% of one DCC and a small proportion of another DCC (Figures 15, 16, 17). This suggests that gene content is being exchanged between the DCCs at different stages during their evolution, and that acquisition of one of these circulating mobile elements acquired by a DCC prior to its expansion, could potentially be contributing to the continuing expansion of a different DCC lineage or contributing to the expansion of a less prevalent lineage.

Whilst there were multiple examples of these clusters, a couple stood out. A cluster of genes present in over 90% of DCC2 isolates and only three DCC1 isolates consisted of multiple antioxidant *aphC/TSA* family genes as well sigma factors, which suggested this gene cluster could be enabling DCC2 to survive under greater oxidative stress as well as quickly respond to environmental cues (Figure 16). However, the lack of clear operons within this cluster coupled with the presence of genes with putative metabolism functions and with roles in cytochrome c biogenesis (Appendix table 1.22) makes the interpretation of the function of this subset of genes within the gene cluster uncertain. Also within this cluster, but within a clear operon, was a alkanesulfonate transport system, which could be helping the DCC to overcome sulfate starvation within the host (261). A gene cluster present in over 90% of DCC1 isolates and seven DCC3 isolates included the *dnd* (Figure 15). The presence of this operon could potentially be providing an advantage to these lineages by enabling them to take advantage of the high concentration of DNA in CF sputum as a nutrient source (82).

The majority of the candidates identified through the pangenome analysis were acquired by only a large proportion of one of the DCCs and there was no overlap between the candidates identified through the SNP density analysis. However, there was evidence to suggest that some of the candidates were participating in similar functions which could suggest that the emergence of the DCC lineages as the most prevalent, whilst not driven by exactly the same changes, were potentially driven by changes affecting the same functional areas (Figure 15, 16, 17).

Gene clusters in DCC1 and DCC2 identified through pangenome analysis and the cluster of candidates identified on the branch leading to the LCA of DCC3 through SNP density analysis all had functions with links to the beta oxidation of fatty acids (Figure 15, Figure 16, Figure 9, Appendix tables 1.21, 1.22, 1.4). Multiple pathogens, including *M. tuberculosis* and *P. aeruginosa*, switch to the degradation of fatty acids as a carbon source within the host (251, 284). Therefore, acquisition of gene content possibly involved in this process suggests that the DCCs may have acquired advantages in this metabolic process over other MABSC lineages. The presence of changes in genes related to this process in all three DCCs suggests it could be the key change that resulted in the increased pathogenic potential of these lineages. However, as is the case for *M. tuberculosis,* it is possible that the MABSC encodes multiple homologs of beta oxidation enzymes, which could be why the DCCs were not found to be statistically significantly enriched with fatty acid metabolism functions (83).

Another area of functional overlap included the presence of two *mspA* porin genes in DCC2 and one *mspA* gene amongst the candidates for DCC3. The msp porin family is the main route by which hydrophilic nutrients are transported across the hydrophobic mycobacterial cell wall (285). Msp porins have been linked both to the resistance to aldehyde disinfectants in *M. chelonae* and *M. smegmatis* and interestingly the number of msp's encoded by an organism has been linked to intracellular persistence, with the deletion of *mspA* and *mspC* in *M. smegmatis* leading to enhanced intracellular survival (286, 287). The role of porins in MABSC organisms has not been investigated and therefore it is unknown whether porins are participating in either resistance to aldehyde disinfectants or possibly contributing to enhanced intracellular survival and thus what role these are potentially playing in the emergence of DCC2 and DCC3.

Enrichment analysis was undertaken to investigate whether the DCCs were statistically significantly enriched with different functions in comparison to the less expanded lineages. For the candidates identified through both SNP density and pangenome analysis no functional enrichment was detected. With regards to the SNP density analysis, this result likely reflected a lack of power as a relatively small number of candidates were identified for each of the DCCs (Table 2). However, in the case of the pangenome analysis, whilst the lack of enrichment could be true, the analysis was also affected by the lack of functional information available for this species. InterProScan was only able to assign GO-terms to 22% of the pangenome candidates present in 1-95% of isolates that are not part of the DCC lineages and to only 77/183, 24/217 and 35/119 of the candidates identified for DCC1, DCC2 and DCC3. Consequently, only a limited picture of the functionality of both the candidates

and the genes encoded in the accessory genome of the non-DCC candidates was captured and this could be affecting the results.

For both the SNP density and pangenome analysis, KEGG pathway analysis produced very few and inconclusive results. Although, in the case of the SNP density analysis the proportions of the groups of pathways detected did mirror the proportion of COG functions assigned to the candidate genes, which did suggest that pathways performing functions in the expected areas were being identified (Figure 11). However, given that only a few candidates from each candidate list were mapped to KEGG pathways and that candidates were often mapped to multiple pathways the results are unclear. Blast2GO identifies the potential pathway a gene is functioning in through using InterProScan to assign the gene to a GO-term, with the GO-term, where possible, assigned an enzyme code and the pathways the enzyme functions in then retrieved from the KEGG pathway database. Therefore, with limited GO-term information and the GO-terms identified not all assigned an enzyme code, missing information hampered this analysis.

Whilst the candidates highlighted through SNP density and pangenome analysis contained genes with potential functional overlap, there were also candidates encoded by only one of the DCC lineages which could potentially provide an advantage to survival in the host and that possibly contributed to the lineage becoming one of the most prevalent in the CF community. On the branch leading to the LCA of DCC1 genes playing a role in homogentisate metabolism and the degradation of phenylacetic acid were identified through pangenome analysis and have both been linked to the virulence potential of the CF pathogens, *P. aeruginosa* and *B. cepacia* respectively, and have previously been described in the reference genome of *M. a. abscessus* ATCC19977 by Ripoll et al. (2009) (82, 179, 267). However, the transport, biosynthesis and metabolism of BCAAs (Figure 15, Appendix table 1.2, 1.21), has not previously been linked with the adaptation of the MABSC to the CF lung.

The cluster of genes associated with the BCKADH complex, was identified through both pangenome and SNP density analysis (Appendix table 1.2, 1.21). This could be due to the reference to which all the isolates were mapped, *M. a. abscessus* ATCC19977, being within DCC1 and thus either this cluster of genes is encoded by all MABSC and the pangenome software Roary split the genes due to either differing flanking genes or the protein sequence identity being less than the selected cutoff for clustering orthologous genes, which in this

analysis was 90% (223). If this was the case, it suggests that potentially the variants of these genes encoded by the early ancestral lineages of DCC1 were advantageous.

Figure 23 shows the BCAA that are metabolized by the BCKADH enzyme complex, and the pathway through which the products of the complex are funneled into the Krebs cycle to provide energy for the organism (244). Genes associated with the transport (MAB_2622c) and biosynthesis of BCAA (MAB_2691), the BCKADH complex enzymes (MAB_0894 - MAB_0897) and potentially genes (MAB_0898-MAB_0918c) which catalyze the breakdown of the products into reactants that are fed into the TCA cycle were all encoded within the SNP density and pangenome candidates identified for DCC1. All these processes have been linked to virulence in other pathogens. Firstly, BCAA transport has been linked to virulence in *Streptococcus pneumonia*, *Staphylococcus aureus*, *Francisella tularensis* and *Legionella pneumophila* (245, 246, 288). Secondly, BCAA biosynthesis mutants in *M. tuberculosis* display attenuated virulence in mice (289). Finally, the disruption of genes involved in catabolism of BCAA in *M. tuberculosis*, including the disruption of *bkdC*, have been shown to result in the lethal buildup of the toxic BCAA intermediates (244). This suggests that the use of BCAA by the DCC1 lineage may be an area worthy of further investigation, and that potentially pathways involved in the transport, biosynthesis and catabolism of BCAA are linked to the pathogenesis of *M. abscessus*.

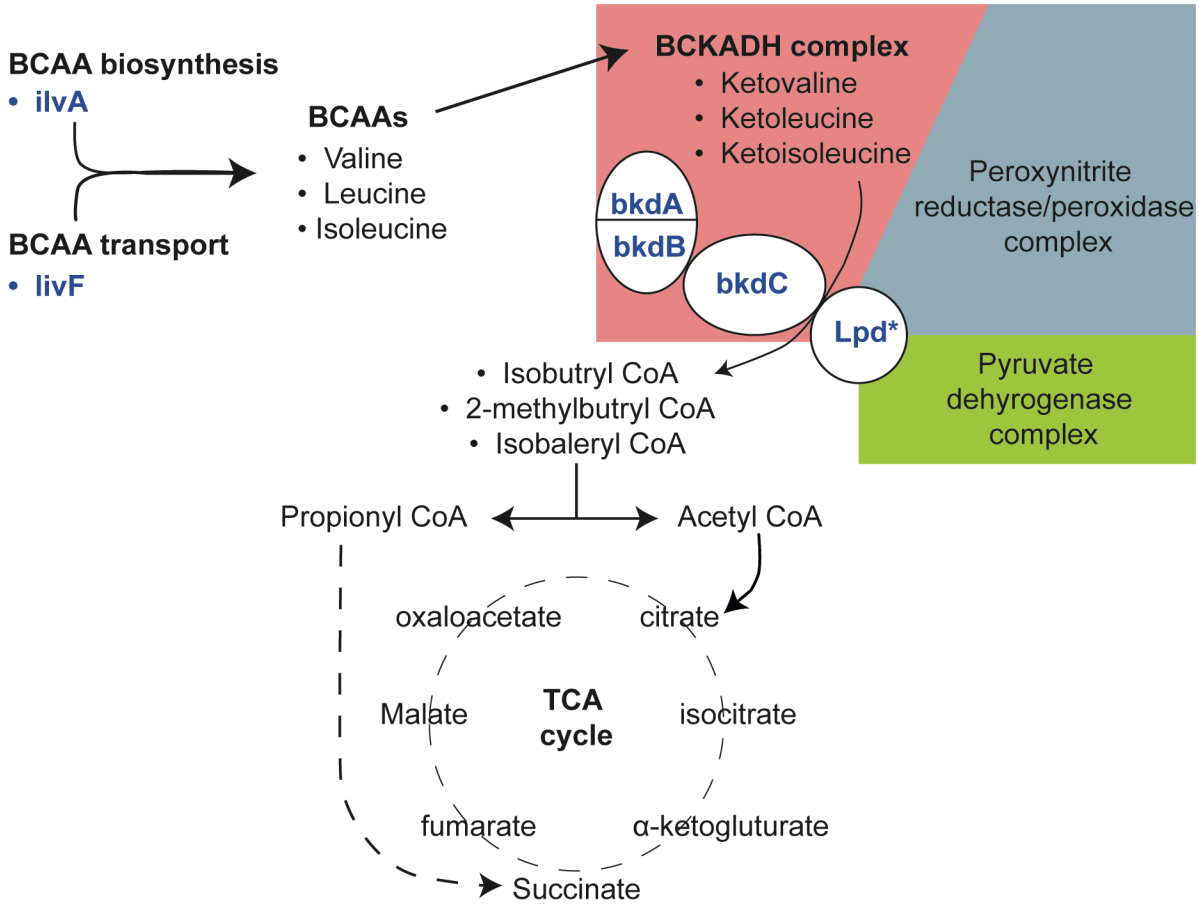


Figure 23: Metabolism of branched chain amino acids potentially playing a role in the emergence of DCC1

Through both SNP density and pangenome analysis genes associated with the biosynthesis, transport and metabolism of branched chain amino acids (BCAA) were identified (blue and bold), which suggested that DCC1 may have been better able to use the abundance of amino acids in the CF sputum as a source of energy. * a candidate was annotated as dihydrolipoamide dehydrogenase (*lpd)*, however, unlike the genes of the BCKADH complex, it was not found to be orthologous to *lpdC* in *M. tuberculosis* H37Rv. Figure adapted from Venugopal et al. 2011 (244).

Further genes with links to virulence amongst the candidates identified as potentially predisposing DCC1 to become the most prevalent included a gene involved in histidine biosynthesis, which is essential for *M. tuberculosis* survival in the host, and the *pstSCAB* operon, which enables the organism to scavenge phosphate from the host in order to be able to keep performing key cellular functions (181, 247, 268, 269). These observations suggest that as well as shared gene content with potential roles in virulence with DCC2 and DCC3 and acquiring gene content with similar functions to that acquired by the other DCCs, LCA of DCC1 had also evolved a unique array of gene content changes and variants in genes with potential roles in virulence and survival within the host.

Contrastingly, the majority of the pangenome candidates identified for DCC2 for which it was possible through sequence homology and literature searches to predict a function were either present in multiple DCCs or functionally overlapped with gene content from other DCCs, and thus have been discussed previously. However, DCC2 did stand out due to the relatively small number of candidates identified through the SNP density analysis, with just six identified. However, the lack of candidates could be due to the fact that three of the six candidates which did incur a significant difference in SNP density were regulators and the subsequent changes in regulation could be as significant phenotypically as changes in SNP density in multiple genes with non-regulatory functions (Figure 9). Without experimental analysis it was not possible to determine which genes were under the control of these regulators, particularly as they didn't clearly form part of an operon. However, one of the regulators, a *tetR* family regulator, MAB_3565, was encoded in close proximity to an *mmpS* gene and *mmpL* gene. *mmpS* and *mmpL* genes are involved in the transport lipids which form part of the highly specialized Mycobacterial cell wall (248). To my knowledge, the functions of these *mmpS* (MAB_3565c) and *mmpL* (MAB_3562c) gene have not been investigated for *M. a. abscessus* ATCC19977, nor were either genes found to be orthologs to any of the *mmpS* or mmpL genes encoded by *M. tuberculosis* H37Rv (230). However, given, the cell wall is a critical virulence determinant in all pathogenic Mycobacteria, if this regulator is playing a role in the regulation of these genes, then it could potentially be a critical change that was associated with the emergence of this lineage as one of the most prevalent in the CF community.

Similarly to DCC1, there was evidence that DCC3 had, in addition to the overlapping gene content with DCC1 and DCC2 respectively and changes in similar pathways to both other DCCs, acquired further unique virulence determinants. Most interestingly, the candidate list included, in contrast to the partial *mce* operons present within the candidates for DCC1 and DCC2, a complete *mce* operon (Figure 17). *Mce* operons encode two *yrbE* genes, which have similarity to membrane spanning proteins, and six *mce* genes. The first *mce* gene in the sequence enables the organism to invade mammalian cells, however the function of the remaining five is unclear, but there is evidence that they are a combination of membrane associated and exported proteins (272). The deletion of the *mce1, mce2* and *mce3* operons in *M. tuberculosis* have been shown to result in attenuated virulence (272). The *mce4* operon enables the import of host cholesterol, a key mechanism by which *M. tuberculosis* is able to persist within the host (290). In terms of gene content the *mce* operon encoded by over 90% of the DCC3 isolates was most similar to *mce1* from *M. tuberculosis* H37Rv, with a *tetR*

regulator, followed by a putative fatty acid degradation gene, two *yrbE* genes and six *mce* genes (177). However, the gene orientation differed and none of the genes were found to be orthologous to the genes within the *mce1* operon encoded by *M. tuberculosis*. However, this could be due to the criteria used to detect orthologous genes using reciprocal blast. Without experimental analysis, it is not possible to determine the function this *mce* operon is performing, however, the rich history of *mce* operons playing a role in the pathogenesis of other pathogenic mycobacteria suggests it could have played a role in the emergence of the DCC3 lineage.

Intriguingly, the vast majority of the pangenome candidates (92%) identified for DCC3 were also encoded by the basal lineage to DCC3, with just 18 candidates not encoded by this lineage. However, the basal lineage was found to be one of the significantly densely clustered lineages with increased pathogenic potential in (73). Thus the candidates identified as possibly linked to the emergence of DCC3 could have also contributed to the emergence of this lineage. The extent to which these two lineages have expanded, however, differs considerably, and DCC3 was also responsible for the outbreaks in CF centers in Papworth in the UK and Seattle in the US as well as an epidemic of SSTIs in Brazil (70, 119, 130). This could suggest that the 18 candidates not encoded by the basal lineage are responsible for the greater prevalence of the DCC3 lineage.

A mobile element encoding a methyltransferase (dpnM) was amongst these 18 candidate genes (Figure 17). The methylome of a bacterial genome can now be detected through SMRT sequencing and this has led to an increased awareness of the impact that a change in methylation pattern can have upon an organism's regulatory circuits and in some cases pathogenicity (214). Thus it was hypothesized that potentially the presence of this methyltransferase could have caused a change in the methylome that led to changes in the regulation of virulence related genes.

As well as the machinery to enable its mobility, the mobile element encoded dpnM and a hypothetical protein with a conserved DGQHR domain. This hypothetical protein was found to have the most convincing sequence similarity to the *dndB* sulfur modification protein (Table 6). However, there was also weak sequence similarity to a YraN endonuclease family protein, and the DGQHR conserved domain has been observed within a type I restriction endonuclease encoded by *Campylobacter jejuni* (291, 292). This is significant because if this second gene is an endonuclease it is likely that the mobile element encodes a restriction modification (RM) system and that protection of host DNA against digestion by the restriction

enzyme would be the main function of dpnM. RM systems protect the host from foreign DNA by recognizing their corresponding methyltransferases unmodified motif in invading DNA and proceeding to cleave the DNA (293). However, the base modifications carried out by methyltransferases that are part of RM systems have also been linked to changes in the transcriptome (293).

Preliminary functional validation of dpnM has begun to shed light on its potential role in the emergence of DCC3. Although, as the KOs generated were of the complete mobile element, whether the hypothetical protein is an endonuclease remains under investigation. Structural analysis, performed by Dr Sony Malhotra, showed that the methyltransferase was homologous to the dpn family of methyltransferases (Figure 21Ai). The introduction of known inactivating mutations resulted in two of the mutations, one in the conserved ado-met binding motif (FXGXG, F42S) and one catalytic domain (DPPY, D184A) being predicted and confirmed to be inactivating and one of the mutations in the catalytic domain (Y187L) predicted and confirmed to be destabilizing (Figure 21 Bi, ii, Ci, ii). This confirmed that the methyltransferase was SAM dependent and, as the presence of the conserved motifs FXGXG and DPPY implied, that the methyltransferase modified the amino group at the sixth position of an adenine (N^6-adenine) (237, 277).

Through SMRT sequencing it was confirmed that dpnM modifies adenine bases at the N^6 position and that the motif recognized by dpnM was either BRGATCC or RGATCC (Table 7). These motifs contained the shorter motif, GATC, that REBASE predicted the dpnM to be modifying. The presence of the IUPAC notations B (either C,G or T) and R (purines), indicated that either the methyltransferase was recognizing an ambiguous base, or that motiffinder was being over-specific and extending the potential motif beyond the true recognition site, a known issue with this program. Furthermore, only 47%, 62% and 73% of the predicted motifs encoded by BIR1049, DEN538 and RHS37 respectively were detected as modified and generally, according to Pacific biosciences, 100% of the predicted motifs encoded by a prokaryote should be modified (294). The fact that several of the attempts to sequence these isolates required a clean-up step before SMRT sequencing was successfully achieved potentially suggests that the quality or amount of the DNA remaining after this step was not adequate enough to produce high enough coverage to detect all the base modifications. The correlation between the percentage of motifs modified in the genome and the average motif coverage (Table 7) suggests that this could be the case. It remains unclear why BIR1049Δinsertion:dpnM failed to restore the BIR1049wt methylome. The complementation plasmid was detected within the *de novo* assembly inferred from the SMRT

sequencing reads. Therefore perhaps the quality of the SMRT sequencing or potentially the loss of function of dpnM on the complementation plasmid could be associated with the failure to reproduce the phenotype. This remains under investigation.

SMRT sequencing also detected the presence of two non-palindromic bipartite modified motifs in the BIR1049wt and BIR1049Δinsertion sequences, which suggested that further methyltransferases were influencing the methylome of the DCC3 lineage. However, further investigation into the methyltransferase genes responsible for this signal was beyond the scope of this analysis.

Most significantly, Daniela Rodriguez-Rincon showed that dpnM was potentially playing a role in the increased intracellular survival associated with the clustered MABSC lineages (Figure 22B) (73). However, the fact that these results were unable to be replicated means that these results are preliminary. Work is ongoing to investigate why it has not been possible to replicate these results. Further work is also in progress to perform differential expression analysis on the BIR1049wt, BIR1049Δinsertion and BIR1049Δinsertion:dpnM isolates, in order identify the genes whose regulation is impacted by dpnM. This will be achieved by correlating the regulatory changes detected with the positions of the modified motifs predicted through SMRT sequencing in the genomes. The functions of the genes under the control of dpnM should then identify in what way it could potentially be contributing to the increased intracellular survival phenotype associated with the DCC3 lineage.

The preliminary functional validation of one of the candidates identified through these analyses suggested that genes which potentially predisposed the DCCs to become the most prevalent in the CF community had been successfully identified through these methods. However, there are areas in which these methods are limited or could be improved. Firstly, the observation that the MABSC infecting population was dominated by significantly densely clustered lineages with increased pathogenic potential was achieved by comparing all significantly densely clustered isolates to those from genetically distinct backgrounds, not comparing the three DCCs to all isolates that were not part of a DCC (73). This was not done initially for the analyses reported here because the extent of the expansion of the DCCs in comparison to the other densely clustered lineages suggested that these lineages were preadapted in such a way that the signal would be evident even in comparison to other lineages with increased pathogenic potential. Therefore, it is possible that candidates associated with all or a large proportion of significantly densely clustered lineages are being

missed. In the future this analysis should be expanded to compare all the densely clustered lineages against the isolates from genetically distinct backgrounds.

Whilst the SNP density analysis did highlight some potentially interesting candidates, two possible limitations should be noted. Firstly, the inability to remove recombination from a phylogeny with such deep branches meant that variants acquired through recombination, as opposed to through the independent acquisition of multiple nonsynonymous SNPs, could have been detected. Secondly, the majority of the candidates identified through the SNP density analysis, were based on the accumulation of very few nonsynonymous SNPs by the gene, both on the branch leading to the LCA of the DCC and on the branches evolving independently to the DCCs. Consequently, a Fisher's exact test as opposed to a X^2 test might have been more appropriate. However, the way in which the data was generated meant it was not possible, in the time remaining, to apply the Fisher's exact test. Pangenome analysis comes with known challenges, such as splitting paralogs correctly and correctly selecting a cutoff for the percent identity expected from the same gene present in genetically distinct lineages. Further analysis is needed to decipher if these issues are affecting these results. Despite these weaknesses, areas of interest for future research where highlighted through this analysis and preliminary functional validation has begun to explain the way in which one of the candidates identified could have provided an advantage that contributed to the emergence of the DCC3 lineage.

2.5 Conclusions and future directions:

Through SNP based and pangenome analyses the aim of this research was to investigate whether the same genetic or functional factors drove the emergence of the three most prevalent MABSC lineages in the CF community. The results showed that there was no single factor that drove the emergence of all three DCCs. Although, there was evidence that the three DCCs had potentially undergone changes in the same functional area, with all three DCCs found to have potentially acquired changes in the genes associated with the beta oxidation of fatty acids in their evolutionary history just prior their LCAs. Whilst the remaining candidates did include examples of the same gene content encoded by two of the three DCCs and the DCCs evolving in similar functional areas, the bigger picture suggested the three DCC lineages had emerged independently from genetically distinct backgrounds. However, the candidates identified for each DCC did include multiple genes that could have played a role in adapting each DCC lineage to be successful within the host, with the preliminary functional follow up analysis on one of the candidates showing that it was

possibly associated with increasing intracellular survival. Given the limited evidence of either SNP based, gene content or functional convergence between the three DCC lineages, it seems unlikely that there is a single genetic state of *M .abscessus* that is preadapted to the human host. Thus, each of the three DCC lineages have evolved to encode a unique combination of virulence determinants that provided an advantage in the CF lung, resulting in their emergence as the most prevalent lineages in the CF community. The fact that this has happened independently on different genetic backgrounds using different gene combinations suggests that the drive for this emergence is a change in the host niche, rather than the pathogen, and the obvious candidate for this is the increased numbers of CF patients surviving for much longer, significantly increasing the available niche for the DCCs to expand into.

Significantly more work is required to understand whether the same genetic or functional factors drove the emergence of the DCC lineages as well as the smaller significantly densely clustered lineages responsible for the majority of MABSC infections in the CF community. The analysis needs to be performed comparing all the significantly densely clustered lineages to the genetically distinct lineages. Experimental and bioinformatics analysis on the most promising candidates, such as the role of the beta oxidation of fatty acid pathways in all the DCCs, the use of BCAA by DCC1 and the ubiquinone biosynthesis operon encoded by DCC1 and DCC2 is required, as is further analysis into the role dpnM is playing in DCC3. As novel tools are developed it may be possible to address this question in a different way and come to more clear conclusions that will enhance the understanding of both how the most prevalent lineages of the MABSC have emerged as well as highlight the key areas in which the MABSC is adapting to become a more efficient opportunistic pathogens.

3. Genetic changes driving the continuing expansion of the recently emerged and more virulent *Mycobacterium abscessus* species complex lineages

Statement of contribution: I performed all the bioinformatics analysis reported in this chapter. The project was developed and supervised by Julian Parkhill and Andres Floto. Julian Parkhill, Andres Floto, Simon Harris and Josephine Bryant contributed to the interpretation of the results.

3. Continuing expansion of the clustered lineages

3.1 Introduction

The MABSC global population structure showed that the majority of MABSC infections in the CF community were caused by a few lineages which had emerged recently and spread widely (73). Molecular phenotyping showed that these recently expanded and widely disseminated lineages (referred to as 'clustered' lineages) were more virulent than isolates from genetically distinct backgrounds which suggested that they had adapted to the CF lung (73). In chapter 2 the genetic changes that occurred prior to the clonal expansion of the three most prevalent MABSC lineages, the DCCs, were examined and it was shown that no single genetic factor had predisposed the DCCs to become the most prevalent in the CF community. The aim of this chapter was to investigate the genetic changes that occurred during the clonal expansion of the recently emerged and expanding lineages. To increase the signal available, given the small number of SNPs observed during the clonal expansion of the clustered lineages and because the observation that these lineages had increased virulence was determined by comparing clustered vs unclustered lineages, this analysis investigated the changes that occurred during the clonal expansion of all the recently emerged clustered lineages (73).

By investigating the changes occurring during the clonal expansion of the clustered lineages and looking for convergence occurring between the lineages as they have expanded and spread amongst the CF community, genes associated with the ongoing adaptation and spread of these lineages could be identified. Given that the clustered lineages have been shown to be more virulent and to be capable of indirect person to person transmission, the genes under positive selection could be associated with these characteristics, which currently are poorly understood (70, 73, 119).

3.2 Methods

3.2.1 Mapping, variant calling and phylogenetic analysis

The single isolate per patient dataset described in section 7.1.1.1 supplemented with 29 publicly available isolates was used in this analysis. Initially, using the mapping, variant calling and phylogenetic methods described in sections 7.3, 7.4 and 7.5, phylogenetic trees for each subspecies were inferred from the variable site alignments generated after mapping the 555 isolates to their subspecies reference genomes (Table 15) in order to increase the mapping resolution whilst maintaining the ability to detect convergence between clustered

lineages. However, in order to increase the power to detect convergence between clustered lineages from different subspecies the 152 *M. a. massiliense* subspecies isolates and 32 *M. a. bolletii* isolates were also mapped to the *M. a. abscessus* ATCC19977 reference genome, using the methods described previously.

3.2.2 Phylogenetic Clustering

TreeGubbins (developed by Simon Harris, for further details see section 7.5) was used to identify significantly dense nodes within the phylogenies (73). All the descendants of a significantly dense node were classified as clustered isolates. Isolates which did not fall within a cluster were classified as unclustered. Only clusters consisting of at least five isolates, to increase power, and with isolates from multiple locations, to rule out all isolates from a cluster that were potentially acquired from the same environmental source, were investigated in this analysis.

3.2.3 Identifying genes under positive selection

By mapping the SNPs back onto each of the phylogenies, using an inhouse script (developed by Simon Harris, for further details see section 7.4), the position and effect (synonymous, nonsynonymous, nonsense, intergenic) of each variant predicted to have occurred between the reconstructed ancestors of each node were identified. The SNPs of interest in this analysis were those that were accumulated i) on the branches after the clonal expansion of each clustered lineage and ii) that had occurred between nodes that had generated multiple progeny (i.e terminal branch SNPs were removed). Terminal branch SNPs were removed as the aim of this analysis was to detect genes under selection during the clonal expansion of the clustered lineages and it could be argued that only for the variants that occur between nodes that go on to produce multiple progeny is there evidence to show that they have contributed to the continuing expansion of the clustered lineages.

To identify genes under positive selection SNPs potentially acquired via recombination were required to be removed. Consequently, for each branch, SNPs were discounted if three or more had occurred within an 1000bp window. The remaining variants were then summed to determine the number of nonsynonymous, nonsense and synonymous SNPs accumulated by each gene on the branches within the clustered lineages. Genes which had accumulated a greater number of nonsynonymous SNPs than would have been expected by chance were determined using a method based upon Ding et al's (2008) 'burden of mutation' approach as described in section 7.11 (295).

3.2.4 Candidate genes follow up analysis

The annotations for candidate genes identified through this analysis were enhanced by searches against the Pfam (v.31.0) and InterPro (v.68.0) databases (224). The orthogroup catalogue developed by McGuire et al. (2012), was used to identify if any of the candidates genes were orthologous to genes encoded by *M. tuberculosis* H37Rv (230). Phage regions within the *M. a. massiliense* CIP108297 and *M. a. bolletii* BD reference genomes were detected using PHASTER (227, 228).

All the isolates, excluding the publicly available isolates, were *de novo* assembled and annotated as described in sections 7.6 and 7.7. An *in silico* PCR approach, applied using an inhouse script (developed by Simon Harris), was used to determine the presence or absence of the operon believed to be controlled by one of the candidate genes in the assemblies of all the isolates. Ten mismatches were permitted between the 81bp primers designed to match the start and end of the region of interest. In order to identify the presence of other proteins in the same family as the candidate genes, the amino acid sequences for all the CDSs of each isolate were extracted and each CDS was searched against the Pfam-A (v.3.1b2) profile database using Hmmer (v.1.1) (296).

In some cases it was useful to perform comparisons either between the three reference genomes, *M. a. abscessus* ATCC19977, *M. a. massiliens*e CIP108297 and *M. a. bolletii* BD or between these reference genomes and *M. tuberculosis* H37Rv, in order to investigate the level of sequence similarity and gene synteny. This was done using tblastx (v.2.2.25) and blastn (v2.4.0), with an e-value threshold of 0.001 and the match length required to be greater than 10. The results were visualized using ACT (v. 13.0.0) and EasyFig (v. 2.1.1) (276, 297).

3.3 Results

3.3.1 Detecting the recently emerged MABSC lineages circulating in the CF community

Phylogenetic trees were constructed for each subspecies after the 556 isolates were mapped to their respective subspecies reference genomes, *M. a. abscessus* ATCC19977, *M. a. massiliense* CIP108297 and *M. a. bolletii* BD. TreeGubbins was used to identify the significantly dense nodes within each of the subspecies phylogenies (Table 8, appendix table 2.1, 2.2, 2.3). The descendants of the significantly dense nodes which were made up of more than five isolates and of isolates from more than one location were hypothesized to represent lineages of the MABSC which had recently emerged and spread amongst the CF community (Figure 24). Within *M. a. abscessus*, *M. a. bolletii* and *M. a. massiliense*, 11, one and eight recently emerged, circulating lineages were identified respectively (Figure 24).

Table 8: Number of clusters detected by TreeGubbins for each subspecies phylogeny

	M. a. abscessus	***M. a. bolletii***	***M. a. massiliense***
Total number of clusters	39	5	14
Total number of clusters of interest	11	1	8
Total number of isolates in clusters	249	5	111
Total number of unclustered isolates	124	27	40

Bryant et al. (2016) showed that the clustered lineages[6] within the MABSC global population were more virulent than the unclustered lineages, which suggested that these lineages had an advantage over other MABSC lineages within the lung or hospital environment (73). The genetic determinants which occurred before the clonal expansion of three largest clustered lineages were investigated in chapter 2. In this chapter the genetic changes that occurred during the clonal expansion of the clustered lineages were investigated to identify genes under positive selection as these genes could have contributed to the dominance of these lineages in the CF community by increasing their virulence and/or transmissibility.

[6] The TreeGubbins clusters determined by Bryant et al. (2016) were regenerated for this analysis. 69/525 isolates used in both analyses were assigned differently in the two analyses (22 were identified as clustered in this analysis but part of clusters not meeting the criteria of interest in Bryant et al's analysis and 47 were found to be unclustered in this analysis and clustered in Bryant et al's analysis). This is likely to be due to using different mapping programs, the addition of the publicly available isolates in this analysis and using different versions of TreeGubbins.

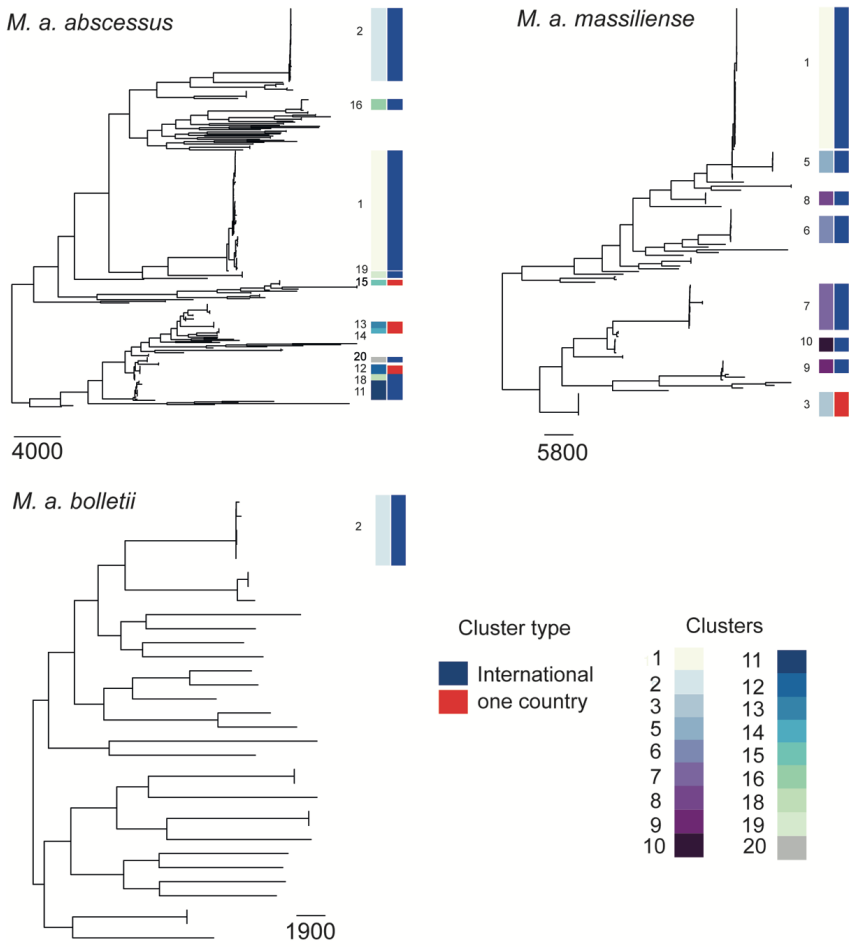


Figure 24: Significantly clustered lineages within the three MASBC subspecies phylogenies when mapped to their subspecies reference genomes

Phylogenetic trees for each subspecies of the MABSC inferred from the variants identified after mapping the isolates to their corresponding reference genomes, *M. a. abscessus* ATCC19977, *M. a. massiliense* CIP108297, *M. a. bolletii* BD. The clusters identified by TreeGubbins are marked, along with whether they consisted of isolates from a single country or international. Clusters which consisted of isolates from only one CF centre or less than five isolates were not included in further analysis. 11 significant clusters were identified within the subspecies *M. a. abscessus*, eight significant clusters were identified within subspecies *M. a. massiliense* and one within subspecies *M. a. bolletii*.

3.3.2 Genes under selection both within clusters and between clusters

Once the SNP dense regions occurring independently on each branch after the LCA of the clustered lineages and the SNPs occurring on the terminal branches were removed, a total of 1,287 SNPs remained on branches within the 11 *M. a. abscessus* clusters, 584 SNPs

remained on the branches within the eight *M. a. massiliense* clusters and 15 SNPs remained on the branches within the one *M. a. bolletii* cluster (Figure 25).

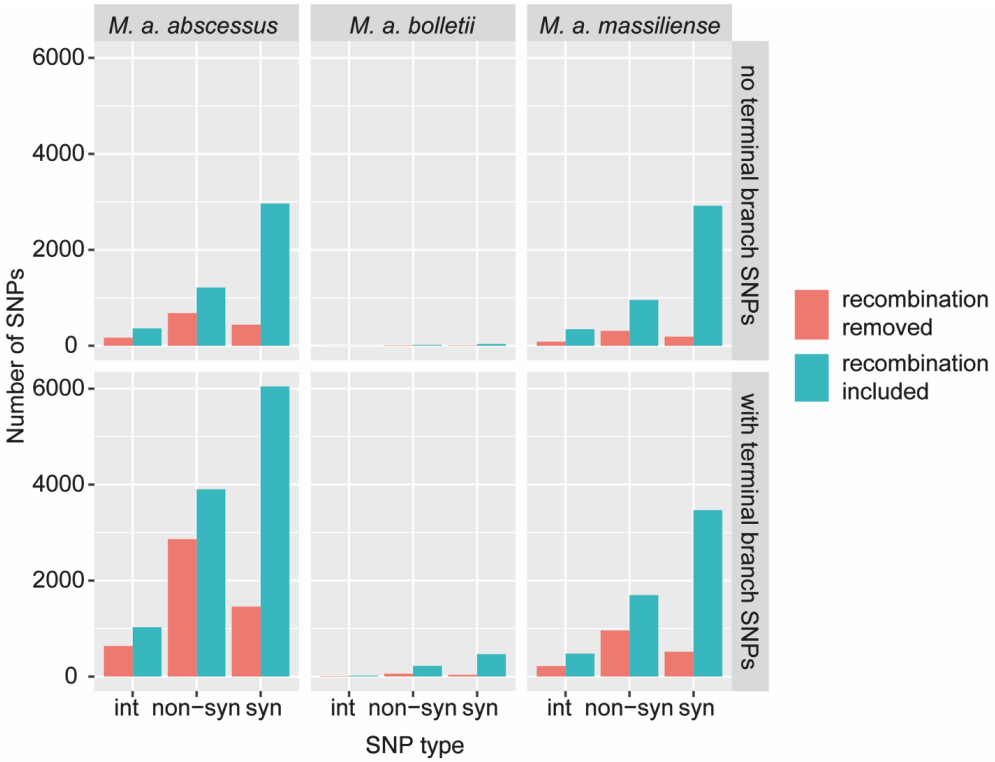


Figure 25: Distribution of SNPs removed due to recombination or due to occurring on terminal branches when the isolates were mapped to their subspecies reference genomes.

Summary of the number and type of SNPs on the branches after the clonal expansion of the clustered lineages when terminal branch SNPs were included (lower panel) and terminal branch SNPs were discounted (upper panel). The colors represent the number of SNPs removed due to potentially be acquired via recombination. This showed the majority of SNPs removed due to recombination were synonymous. These SNPs were based on the phylogenies inferred from the alignments generated after mapping all the isolates were mapped to their respective subspecies reference genomes.

Within *M. a. abscessus* clustered lineages, 560 genes accumulated a nonsynonymous SNP, with the maximum number of nonsynonymous SNPs accumulated by a single gene being eight (Appendix table 2.4). At least one nonsynonymous SNP was gained by 227 genes within *M. a. massiliense* clustered lineages, with the maximum acquired by a single gene being four (Appendix table 2.6). After the clonal expansion of the *M. a. bolletii* clustered lineage the maximum number of nonsynonymous SNPs accumulated by a gene was one, with only eight genes accumulating at least one nonsynonymous SNP (Appendix table 2.5).

In total three genes were found to have accumulated a greater number of nonsynonymous SNPs than would have been expected by chance using a single tailed binomial test ($P < 0.01$). Two genes were found to be under positive selection after the clonal expansion of *M. a. abscessus* clustered lineages, one gene after the clonal expansion of *M. a. massiliense* clustered lineages and no genes were found to have accumulated a significant number of nonsynonymous SNPs after the clonal expansion of the one *M. a. bolletii* clustered lineage (Table 9, for the full results see appendix table 2.7, 2.8 and 2.9).

MAB_4027 accumulated seven nonsynonymous SNPs on the branches after the clonal expansion of *M. a. abscessus* cluster 1 (Table 9). Contrastingly, the five nonsynonymous SNPs accumulated by MAB_2292c were acquired on the branches after the clonal expansion of both *M. a. abscessus* cluster 1 and cluster 16 and similarly the two nonsynonymous SNPs accumulated by CIP108297_03869 (MAB_3906c) occurred on the branches after the clonal expansion of *M. a. massiliense* clusters 2 and 8 (Table 9). This provided evidence that suggested that clustered lineages from differing genetic backgrounds were evolving in the same way.

Table 9: Genes under selection after the clonal expansion of the clustered lineages when the isolates were mapped to their subspecies reference genomes

Subsp.	Locus	Observed	expected	P-value	Clusters	Product	Pfam	InterPro	*M. tuberculosis* H37Rv ortholog
M. a. abscessus	MAB_4027	7	0.185	$1.35x10^{-11}$	1	*tetR* regulator	Helix turn helix DNA binding domain	Helix turn helix DNA binding domain, Tetracyclin repressor C-terminal domain	NA
M. a. abscessus	MAB_2292c	5	0.314	0.002	1, 16	Hypothetical protein	NA	NA	NA
M. a. massiliense	CIP108297_03869 (MAB_3906c)	2	0.013	0.002	6, 8	Hypothetical protein	NA	NA	NA

All the genes that accumulated a significant number of nonsynonymous SNPs on the branches after the clonal expansion of the clustered lineages were present in each of the subspecies reference genomes. However, whilst there was evidence of overlap between different clustered lineages of the same subspecies, there was no overlap between the significant genes identified between the subspecies and nor were nonsynonymous SNPs accumulated by these genes on the branches after the clonal expansion of the lineages in other subspecies. Consequently, given the small number of candidate genes identified, the isolates were all mapped to the *M. a. abscessu*s ATCC19977 reference genome in an attempt to increase the power to detect changes between clustered lineages in different subspecies.

3.3.3 Parallel evolution not detected between clustered lineages from different subspecies

TreeGubbins was used to identify the significantly clustered lineages from the phylogenies constructed from the variant positions extracted from the alignments produced from mapping the 152 *M. a. massiliense* and 32 *M. a. bolletii* isolates to *M. a. abscessus* ATCC19977 reference genome. One clustered lineage consisting of greater than five isolates and isolates from more than one location was identified within the *M. a. bolletii* subspecies phylogeny, whilst nine were identified within the *M. a. massiliense* subspecies phylogeny (Table 10, Figure 26, appendix table 2.2, 2.3).

Table 10: TreeGubbins clusters detected within the phylogenies for each subspecies when all isolates were mapped to *M. a. abscessus* ATCC19977

	M. a. abscessus	***M. a.bolletii***	***M. a. massiliense***
Number of clusters	39	5	11
Number of clusters of interest	11	1	9
Number of isolates in clusters	249	5	113
Number of unclustered isolates	124	27	38

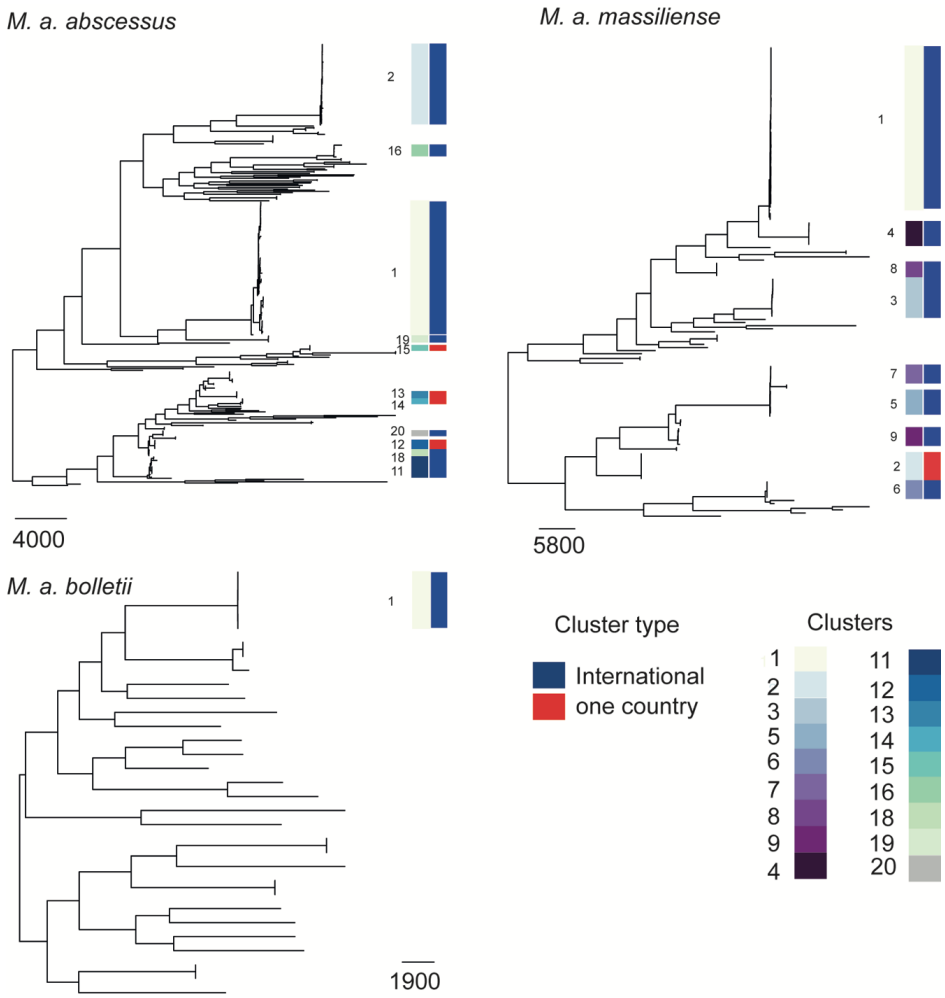


Figure 26:Phylogenetic trees for each subspecies of the MABSC inferred from the variants identified after mapping all the isolates to the *M. a. abscessus* ATCC19977 reference genome.

The metadata columns represent the TreeGubbins clusters and whether they consist of isolates from a single country or international. Clusters which consisted of isolates from only one CF centre or less than 5 isolates were not included in further analysis. 11 clusters were identified within *M. a. abscessus*, nine within subspecies *M. a. massiliense* and one within subspecies *M. a. bolletii*.

In total 623 SNPs were accumulated on the branches after the clonal expansion of the nine *M. a. massiliense* clustered lineages when SNP dense regions and terminal branch SNPs were discounted (Figure 27) (appendix table 2.10). Only 11 SNPs were identified on the branches after the clonal expansion of the *M. a. bolletii* cluster when terminal branch SNPs and SNP dense regions were removed (Figure 27) (Appendix table 2.10). These were combined with the 1287 SNPs observed on the branches after the clonal expansion of the *M. a. abscessus* clustered lineages. Nonsynonymous SNPs were acquired by 820 genes, with the maximum number of nonsynonymous SNPs gained by a gene being 10 (appendix table 2.10, 2.11).

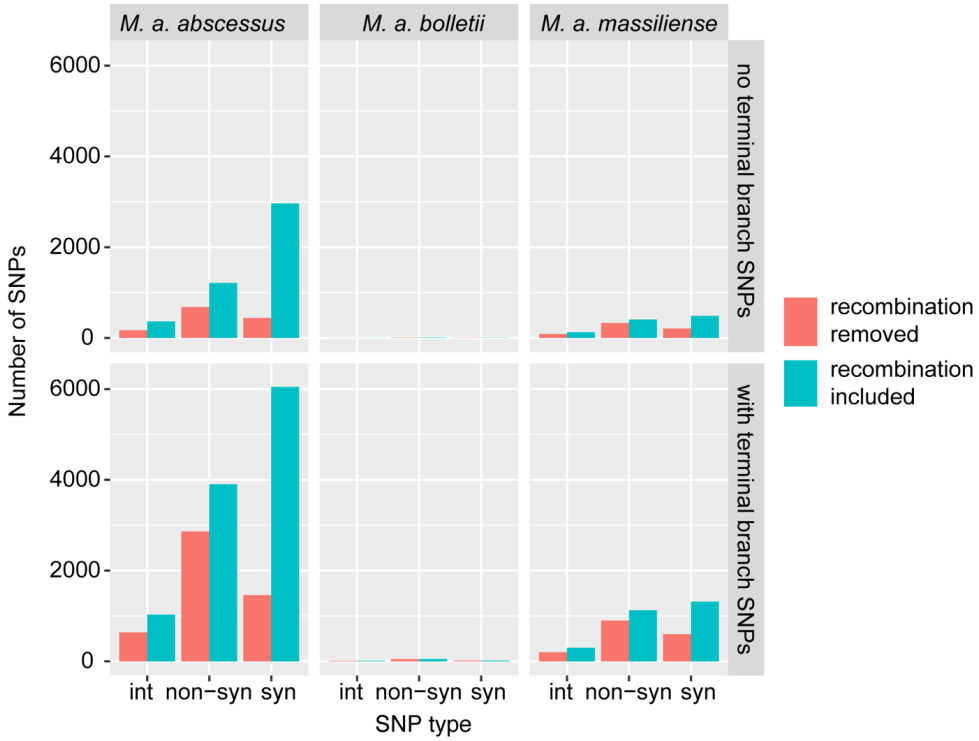


Figure 27: Bar charts summarizing the impact on the SNP count caused by removing recombination and removing SNPs on terminal branches.

Summary of the number and type of SNPs on the branches after the clonal expansion of the clustered lineages when terminal branch SNPs were included (lower panel) and terminal branch SNPs were discounted (upper panel). The colors represent the number of SNPs removed due to potentially be acquired via recombination. This showed the majority of SNPs removed due to recombination were synonymous. These SNPs were based on the phylogenies inferred from the alignments generated after mapping all the isolates were mapped to the *M. a. abscessus* ATCC1977 reference genome.

A single tail binomial test found only one gene had accumulated more nonsynonymous SNPs than would have been expected by chance ($P < 0.01$) (appendix table 2.11). Similarly to the previous analysis, MAB_4027 accumulated a significant number of nonsynonymous SNPs, seven, all within *M. a. abscessus* cluster 1 (p-value $2.96X10^{-06}$, appendix table 2.11). Given that no new candidates were identified through mapping the isolates to the same reference genome, the three candidates identified from mapping the isolates to the subspecies reference genomes were analysed further.

3.3.4 Regulator of a mammalian cell entry operon under selection in M. a. abscessus cluster 1

The only candidate gene identified for *M. a. massiliense* clustered lineages was CIP108297_03869 which accumulated two nonsynonymous SNPs on branches within *M. a. massiliense* clusters 6 and 8 (Figure 28), with the two branches having three descendants each. CIP108297_03869 consists of 123 nucleotides (41 amino acids) and encoded a hypothetical protein. No conserved domains were detected within the gene when it was compared against the Pfam and InterPro databases. CIP108297_03869 was flanked upstream by a hypothetical protein and downstream by a luxR family regulator.

Given the short length of the gene and lack of conserved domains, the possibility that the gene had been disrupted by a result of a miss-assembly or by a contig break was investigated. The 88 contigs that made up the *M. a. massiliense* CIP108297 reference genome were re-ordered to follow the order of the *M. a. abscessu*s ATCC19977 reference genome for this analysis and the re-ordered contigs were re-annotated using Prokka. Whilst CIP108297_03896 was found not to be annotated as a CDS in the original *M. a. massiliense* CIP108297 reference genome, nor did a contig break occur where CIP108297_03896 was predicted to be encoded by Prokka. Interestingly, a CDS was predicted at this loci within *M. a. abscessus* ATCC19977 (Appendix Figure 2.1). However, the equivalent gene in *M. a. abscessus* ATCC19977, MAB_3906c, was found to be significantly longer, consisting of 77 as opposed to 41 amino acids. By looking at the available transcription start sites, codon specific GC content and ribosomal binding sites the start codon of MAB_3906c was determined to be accurate and thus the start of the gene in *M. a. massiliense* CIP108297_03896 appeared to have been deleted. Using the coverage data of the 152 *M. a. massiliense* isolates mapped to *M. a. abscessus* ATCC19977, the start of MAB_3906c was found to have been deleted in 133 of the 152 isolates. The 17 isolates that make up *M. a. massiliense* cluster 7 and 2 unclustered isolates encoded a full length version of this gene.

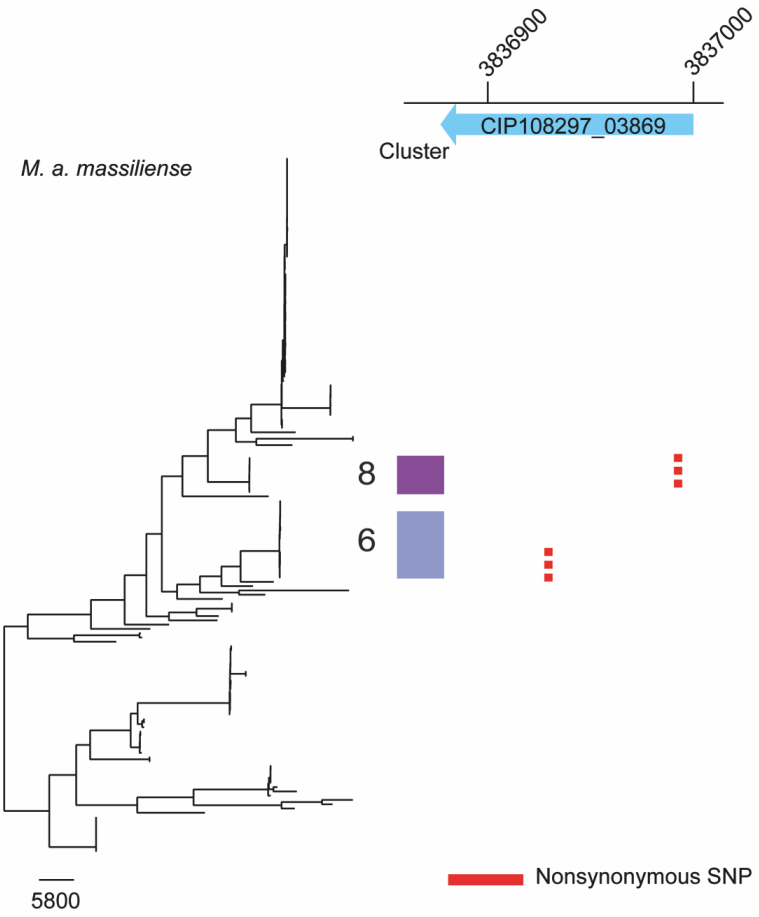


Figure 28: Distribution of the nonsynonymous SNPs accumulated by CIP108297_03869 across the *M. a. massiliense* subspecies phylogeny

The metadata column indicates the two clustered lineages, 6 and 8, that accumulated five nonsynonymous SNPs (red lines) within CIP108297_03869 during their clonal expansion. The phylogeny and clusters used in this figure were those generated after mapping *M. a. massiliense* isolates to *M. a. massilinese* CIP108297.

The remaining two candidates both accumulated a significant number of nonsynonymous SNPs after the clonal expansion of *M. a. abscessus* clustered lineages. MAB_2292c accumulated five nonsynonymous SNPs, four on branches within *M. a. abscessus* cluster 1 and one on a branch within *M. a. abscessus* cluster 16 (Figure 29). These branches had 10 descendant isolates. MAB_2292c, encodes a hypothetical protein, with no conserved domains detected when the amino acid sequence was compared against the Pfam and InterPro databases, whist it was also not found to have an ortholog within *M. tuberculosis* H37Rv (230). The gene is 1121 nucleotides in length (374 amino acids) and is flanked upstream by a hypothetical protein and downstream by a MarR family regulator (Figure 29).

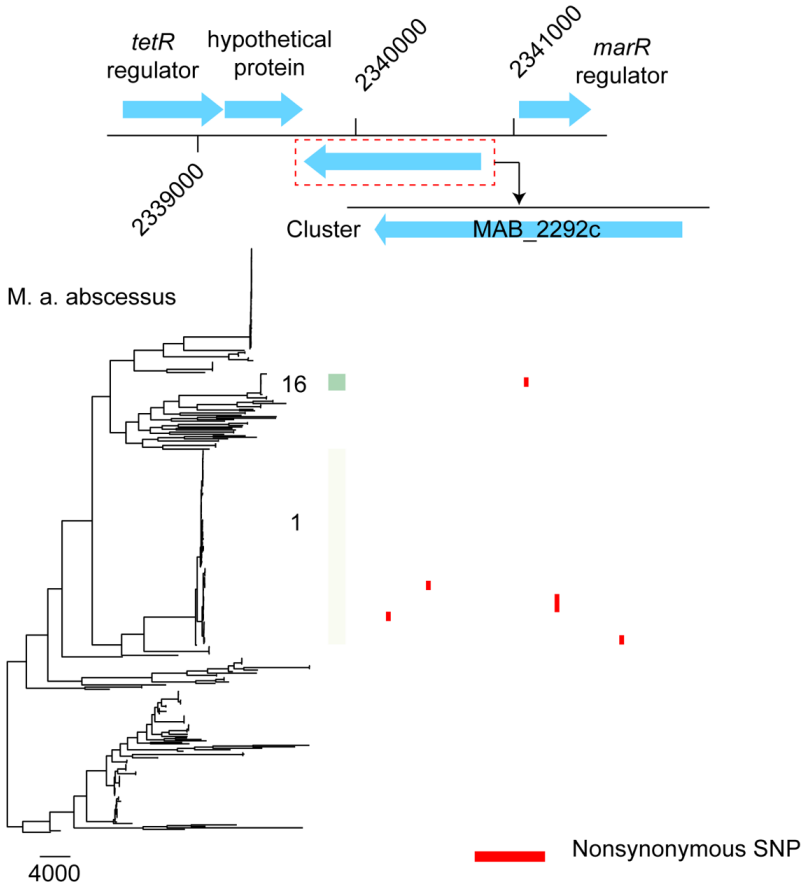


Figure 29: Distribution of the five nonsynonymous SNPs accumulated within MAB_2292c

Five nonsynonymous SNPs were accumulated within MAB_2292c during the clonal expansion of *M. a. abscessus* cluster 1 (DCC1) and *M. a. abscessus* cluster 16. The function of MAB_2292c is unclear with no conserved domains detected within the gene. MAB_2292c was flanked by a *tetR* regulator, a hypothetical protein and *marR* regulator. These failed to provide further insight into the function of MAB_2292c.

MAB_4027 accumulated seven nonsynonymous SNPs on seven branches within *M. a. abscessus* cluster 1, with 36 descendant isolates (Figure 30). MAB_4027 encodes a *tetR* regulator, with the searches of the amino acid sequence against the Pfam and InterPro databases showing a helix-turn-helix DNA binding domain was encoded by amino acids 27 to 73 and a C-terminal ligand binding domain was encoded by amino acids 102 to 183. Five of the seven nonsynonymous SNPs fell within the C-terminal ligand binding domain. MAB_4027 was, similarly to MAB_2292c, not found to be orthologous to any genes in *M. tuberculosis* H37Rv (230). Directly downstream of MAB_4027 two *yrbE* family genes were encoded (MAB_4028, MAB_4029) followed by six *mce* genes (MAB_4030-MAB_4035) and two hypothetical proteins (MAB_4036, MAB_3037) (Figure 30). No conserved domains were

detected to be encoded by the two hypothetical proteins when they were compared against the Pfam and InterPro databases. This sequence of genes is consistent with that of an *mce* operon, which are key virulence factors in *M. tuberculosis* H37Rv (259).

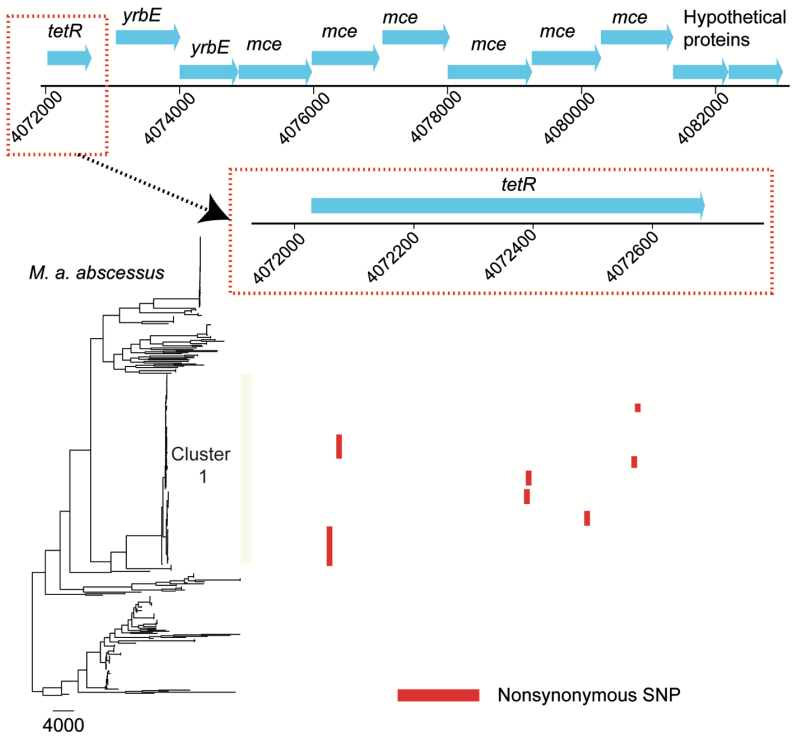


Figure 30: A regulator of an *mce* operon accumulated a significant number of nonsynonymous SNPs during the clonal expansion of *M. a. abscessus* cluster 1 (DCC1)

The maximum likelihood phylogenetic tree of the *M. a. abscessus* subspecies with the metadata aligned representing the seven SNPs accumulated by the *tetR* regulator encoded by MAB_4027. The *tetR* family regulator was encoded directly upstream of a complete *mce* operon. Five of the seven nonsynonymous SNPs were acquired within the C-terminal ligand domain, which was predicted to be encoded by amino acids 102-183.

As MAB_4027 could potentially be regulating this *mce* operon, the operon was compared against the four well characterized and virulence-associated *mce* operons encoded by *M. tuberculosis* H37Rv. The tblastx comparisons between the four *mce* operons and the *mce* operon being investigated here showed that none of the *mce* operons shared complete synteny (Figure 31) and that, whilst there was some sequence conservation between the *yrbE* genes, the level of AAI between the *mce* proteins was minimal and there was no AAI between the putative regulator (MAB_4027) of the candidate *mce* operon and the regulators of the three *M. tuberculosis* H37Rv *mce* operons which encode a regulator directly upstream

of the eight conserved *mce* operon genes. In fact only the *mce4* operon encoded by *M. tuberculosis* H37Rv was orthologous to an *mce* operon encoded by *M. a. abscessus* ATCC19977 (Appendix Figure 2.2).

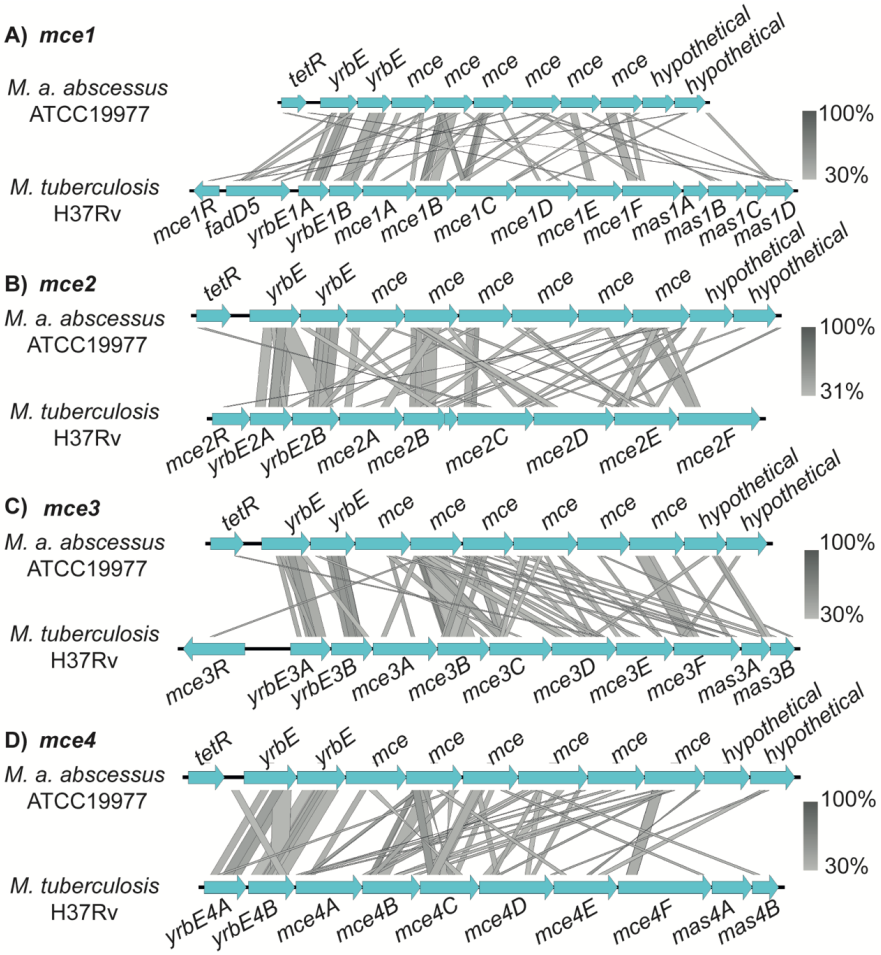


Figure 31:*mce* operon under the control of MAB_4027 does not share complete gene synteny with any of the four *mce* operons encoded by *M. tuberculosis* H37Rv

Tblastx comparisons between the *mce1* (A), *mce2* (B), *mce3* (C) and *mce4* (D) operons encoded by *M. tuberculosis* H37Rv and the *mce* operon under the control of the *tetR* regulator, MAB_4027. The maximum e-value permitted was 0.001, the match length had to be at least 25bps and have at least 30% identity. No homology was detected between MAB_4027 and the regulators of the *mce* operons in *M. tuberculosis* and nor did the *mce* operon share complete gene synteny with any of the four *mce* operons encoded by *M. tuberculosis*.

Using an *in silico* PCR approach, the presence and absence of this *mce* operon across the MABSC was investigated to see whether it was unique to *M. a. abscessus* cluster 1, although this was unlikely given the regulator was present in each of the subspecies reference genomes. The *mce* operon was found to be encoded by 505 (96%) of the 525

isolates[7] (Figure 32). The *mce* operon was lost ancestrally once before the clonal expansion of *M. a. massiliense* clusters 5 and 7 and sporadically in seven further isolates.

3.3.5 MABSC clustered and unclustered lineages on average encode the same number of mce operons

The number of *mce* operons encoded by species, including in a previous analysis of the MABSC, has also been associated with virulence, although this is debated (72, 259, 298). Therefore, given that clustered lineages were found to be more virulent than unclustered lineages and that a regulator of an *mce* operon was found to be under selection during the clonal expansion of *M. a. abscessus* cluster 1 (DCC1), the number of *mce* operons encoded by all the isolates in the MABSC was investigated to see if the clustered lineages encoded a significantly different number of *mce* operons to unclustered lineages (73).

To identify whether a different number of *mce* operons was encoded by the clustered and unclustered lineages, a hmmer search for the conserved domains (yrbE: PF02405, *mce*: PF02470 and PF11887) characteristic of *mce* operon genes was performed. An *mce* operon was classified as the presence of two *yrbE* domain containing genes followed by six *mce* domain domaining genes, with one gap permitted. This showed that on average each isolate encoded 6.01 *mce* operons, with clustered lineages encoding on average 6.23 *mce* operons and unclustered isolates encoding on average 6.06 *mce* operons, suggesting that there was no significant difference between the number of *mce* operons encoded by clustered and unclustered lineages[8] (Figure 32).

3.3.6 SNP dense regions are mainly associated with mobile elements

Only three genes were identified as under positive selection on the branches after the clonal expansion of the clustered lineages. In order to identify whether genes had gained a significant number of nonsynonymous SNPs the SNPs acquired via recombination had to be removed. However, these regions could also be contributing to the expansion of these lineages by increasing their virulence and/or transmissibility.

[7] The publicly available isolates were not included in this part of the analysis.
[8] These were based on the clusters identified after all the isolates were mapped to the *M. a. abscessus* ATCC19977 reference genome.

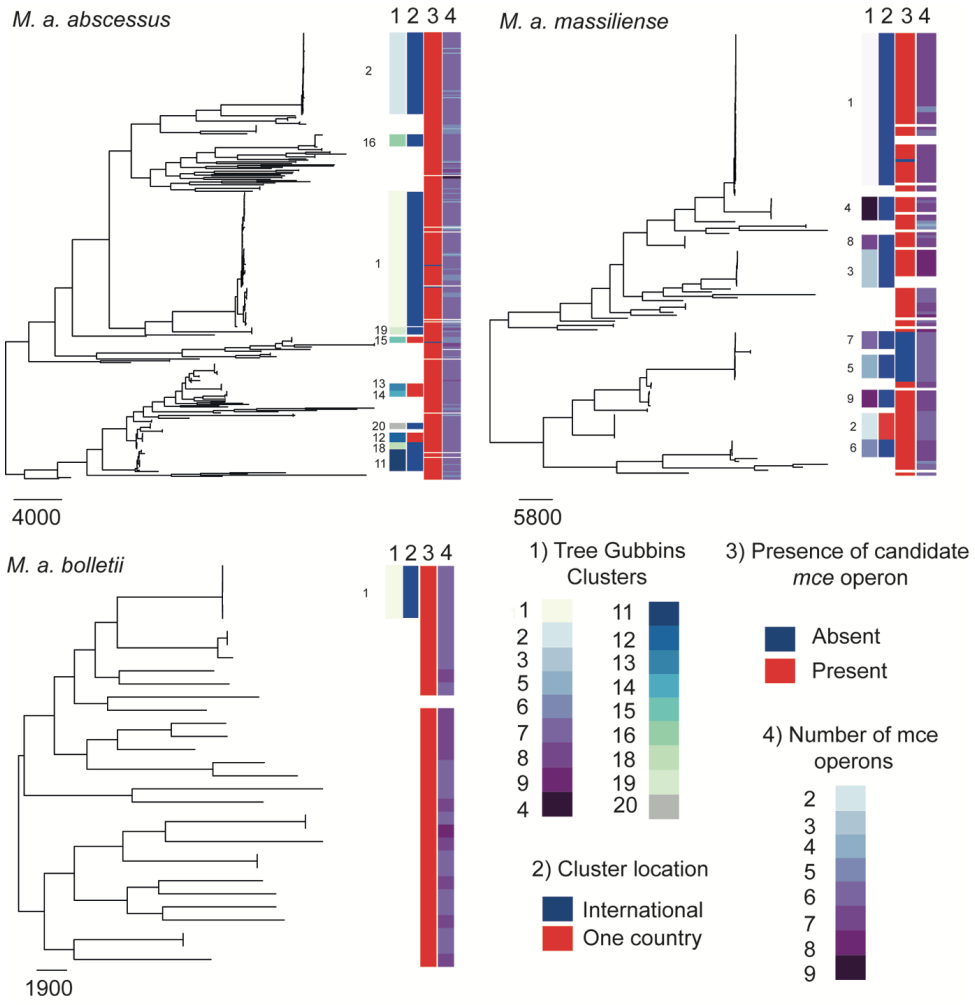


Figure 32: No significant difference between the number of *mce* operons encoded by clustered and unclustered lineages

The metadata aligned to the midpoint rooted maximum likelihood phyogenies of each subspecies represent 1) the TreeGubbins clusters, 2) whether the clusters have spread within a country or internationally, 3) the presence or absence of *mce* operon regulated by MAB_4027 and 4) the number of *mce* operons predicted to be encoded by each isolate. The whites spaces in the 3 and 4 metadata column represent the publicly available isolates that were not included in this part of the analysis.

When the isolates were mapped to their respective reference genomes and terminal branch SNPs were discounted, 186 SNP dense regions consisting of 3,251 SNPs were identified after the clonal expansion of *M. a. abscessus* clustered lineages. 104 SNP dense regions consisting of 3,269 SNPs were identified after the clonal expansion of the *M. a. massiliense* clustered lineages and 6 SNP dense regions incorporating 19 SNPs were identified after the clonal expansion of the *M. a. bolletii* cluster. Significantly more synonymous SNPs were

removed as being due to recombination than nonsynonymous or intergenic SNPs, with 77.6%, 75.1% and 73.1% of the SNPs removed from *M. a. abscessus* clustered lineages, *M. a. massiliense* clustered lineages and *M. a. bolletii* clustered lineages respectively being synonymous (Figure 25, 27).

To identify whether the majority of these SNPs were acquired in known mobile regions, as would be expected, PHASTER was used to identify the potential phage regions within *M. a. massiliense* CIP108297 reference genome and *M. a. bolletii* BD reference genomes (227, 228). Three possible phage regions were predicted within *M. a. massiliense* CIP108297, two complete and one incomplete were detected, whilst one phage region with questionable completeness was detected within *M. a. bolletii* BD (Table 11). The mobile regions encoded by *M. a. abscessus* ATCC19977 have been described previously (82).

Table 11: Phage regions detected by PHASTER

Subspecies	Region length	Region	completeness	score
M. a. massiliense CIP108297	62.4Kb	1789906-1852400	intact	150
M. a. massiliense CIP108297	18.3Kb	3654520-3672829	incomplete	60
M. a. massiliense CIP108297	44.6Kb	3875981-3920589	intact	100
M. a. bolletii BD	71.1Kb	4299517-4370677	questionable	80

For each subspecies, 85%, 72%, and 100% of the genes encoded within the SNP dense regions predicted for *M. a. abscessus*, *M. a. massiliense* and *M. a. bolletii* respectively were associated with mobile elements such as phage or integrated plasmids (Figure 33, A, B) (appendix table 2.12, 2.13, 2.14) (82). Similarly when all the isolates were mapped to *M. a. abscessus* ATCC19977, a combined total of 237 SNP dense regions were detected, incorporating 3,646 SNPs of which 76.6% were synonymous. 82.6% of the SNP dense regions (Figure 33, C) were associated with phage or other mobile genetic elements described in (82) (appendix table 2.14).

Synonymous SNPs also dominated amongst the remaining regions not associated with phage or other mobile elements, with 53% of the 104 SNPs remaining in *M. a. abscessus* clustered lineages being synonymous and 76.6% of 2,011 SNPs remaining in *M. a.*

massiliense clustered lineages. When all the isolates were mapped to *M. a. abscessus* ATCC19977 a similar pattern was observed with 67.9% of the 274 SNPs not associated with predicted mobile elements being synonymous.

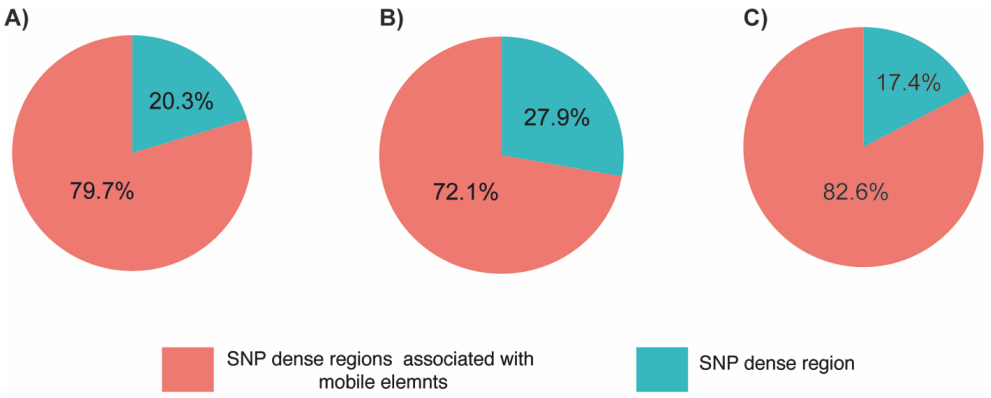


Figure 33: The majority of SNPs removed due to potentially being acquired via recombination are associated with mobile elements

Proportion of the SNP dense regions that were associated with mobile elements when A) all *M. a. abscessus* isolates mapped to *M. a. abscessus* ATCC1997 B) all *M. a. massiliense* isolates mapped to *M. a. massiliense* CIP108297 and C) all the isolates were mapped to *M. a. abscessus* ATCC19977. A figure was not included for *M. a. bolletii* as 100% of the SNP dense regions identified were associated with mobile elements.

3.4 Discussion

The MABSC global population structure consists of multiple clades of significantly densely clustered isolates, suggestive of recently emerged lineages that have spread rapidly to multiple locations (73). These lineages are responsible for the majority of infections caused by the MABSC in people with CF. Molecular phenotyping showed that these lineages are more virulent than unclustered isolates, which led to the hypothesis that these clustered lineages had undergone multiple rounds of evolution within CF lung, enabling them to adapt to become successful lung pathogens and thrive as a CF pathogen (73). In this chapter, the aim was to investigate which genes were under positive selection after the clonal expansion of the clustered lineages, based on identifying SNPs that had occurred in the recent ancestral history of the clustered lineages and which had generated multiple progeny, indicating the continuing success of the lineage. The continuing expansion of the clustered lineages is hypothesized to be being driven by factors such as virulence and/or transmissibility.

Three candidate genes, involving SNPs accumulated after the clonal expansion of 4 lineages, were identified once SNPs on the terminal branches or gained via recombination events were removed (Table 9, Figure 28, 29, 30). For two of the three candidate genes, MAB_2292c and CIP108297_03869, nonsynonymous SNPs were accumulated in more than one cluster, suggesting that clustered lineages from genetically distinct backgrounds are evolving in the same way (Figure 28, 29). Contrastingly, MAB_4027, only accumulated SNPs on the branches after the clonal expansion of *M. a. abscessus* cluster 1, which could imply that the effect of the mutations acquired by this gene is only beneficial in the context of the *M. a. abscessus* cluster 1 genetic background. No convergence was seen between clustered lineages from different subspecies, even when the power to detect core genome SNPs was increased by mapping all the isolates to the same reference genome. However, given only one candidate was identified for *M. a. massiliense* clustered lineages and none were identified for *M. a. bolletii* clustered lineages and that the maximum number of nonsynonymous SNPs acquired by a single gene were four and one respectively in these subspecies, it suggests that a lack of signal is responsible as opposed to this being evidence that lineages from different subspecies of the MABSC are adapting in different ways to the CF environment.

Of the three candidate genes only one, MAB_4027, had an assigned annotation, with no conserved domains detected within either of the other two candidate genes, MAB_2292c and CIP108297_03869 (Table 9). The lack of any conserved domains in these genes meant it was not possible to determine the function of these genes without experimental analysis. The genes flanking MAB_2292c and CIP108297_03869 also failed to shed light on the function of these genes, with both genes flanked downstream by a regulator and upstream by hypothetical proteins and not forming part of a clearly defined operon from which it may have been possible to infer a potential function (Figure 28, 29).

Furthermore, in the case of CIP108297_03869 it is possible that this gene has already lost its function. A blastn comparison showed that CIP108297_03869 had significant sequence homology to the end of MAB_3906c encoded by *M. a. abscessus* ATCC19977, suggesting that the start of CIP108297_03869 had been deleted (appendix figure 2.1). Analysis of read coverage confirmed the deletion, with the loss of the start of this gene occurring in 133 of the 152 *M. a. massiliense* isolates. All the isolates within *M. a. massiliense* cluster 7[9] (n=17), as

[9] Based on the clusters identified by TreeGubbins when all the isolates were mapped to their own subspecies reference genomes.

well as two unclustered isolates, encoded a full length version of this gene. This shows that the majority of *M. a. massiliense* isolates encode a shorter version of this gene, however, it remains unclear firstly whether the shorter version of the gene is functional and if so what impact the nonsynonymous SNPs would have. It should also be noted that given the short length of the gene and the small number of nonsynonymous SNPs accumulated, this candidate gene could also be a false positive.

MAB_4027 accumulated the greatest number of nonsynonymous SNPs involving the most progeny. This gene encodes a TetR repressor and is located upstream of an *mce* operon which it is likely to be regulating (Figure 30). *Mce* operons are known to be critical to the virulence of *M. tuberculosis,* with the four *mce* operons encoded by this organism associated with its ability to invade mammalian cells, prevent phagosome maturation and use cholesterol as a sole carbon source (272, 290, 299, 300). However, tblastx comparisons showed that there was no orthology between the four *mce* operons encoded by *M. tuberculosis* H37Rv and the *mce* operon possibly under the control of MAB_4027 (Figure 31). Consequently, it is not possible to definitively say that this *mce* operon is performing a similar virulence associated function in *M. abscessus* as the *mce* operons encoded by *M. tuberculosis*. However, homologs of *mce* operons are present in all mycobacteria as well as five other genera of actinomycetes, several species of which inhabit an environment much more similar to that of the natural niche inhabited by the MABSC (259). In other soil dwelling actinomycetes *mce* operons have been associated with the import of sterols, virulence against amoebae, changes in biofilm formation and colonization of plant roots (301, 302). Thus, it seems more likely that this *mce* operon is performing a function beneficial in the MABSC's original niche, potentially similar to those described for other soil dwelling actinomycetes, and that a change in regulation of this function is contributing to the ongoing success of *M. a. abscessus* cluster 1.

Three of the four *M. tuberculosis mce* operons encode repressors directly upstream of their corresponding *mce* operons which provides further support to suggest that MAB_4027 is regulating the *mce* operon located just downstream of it (177, 300, 303) (Figure 31). The distribution of the nonsynonymous SNPs accumulated by MAB_4027, with five of the seven occurring in the ligand binding domain, suggests that changes in ligand binding affinity are being selected for, although it is not possible to determine without experimental analysis whether these changes result in increased or decreased expression of the operon. There have been no reports in the literature, to my knowledge, of mutations occurring naturally in the repressors of *mce* operons, with analyses into the function of the regulators of *mce*

operons being carried out using artificial disruption of these genes (300, 303, 304). However, naturally occurring mutations in *mce* genes were identified by Tettelin et al. (2014) when they compared representatives of the Seattle, Papworth and Brazilian (not CF associated) outbreak lineages to isolates not associated with outbreaks (91). Representatives from these outbreaks fall within or close to *M. a. massiliense* cluster 1 (DCC3) which potentially suggests that *mce* operons are playing a role in the adaptation of *M. a. massiliense* cluster 1, although, of the 17 SNPs they observed in *mce* genes only two were nonsynonymous which potentially explains why these were not detected in this analysis (91). Interestingly, *mce* genes have also been associated with the adaptation of the modern *Beijing* strains of *M. tuberculosis,* which have been shown to be more virulent than their ancestral sub lineage and have also spread globally (305).

Given that this analysis showed that the change in regulation of an *mce* operon was potentially associated with the success of one of the most prevalent MABSC lineages and that the number of *mce* operons has previously been noted to differ between MABSC lineages and that differences in the number of *mce* operons encoded by species of actinomycetes has in the past been tentatively associated with virulence, the number of *mce* operons encoded by the clustered and unclustered lineages were compared (72, 298). However, no difference was observed between the number of *mce* operons encoded by the clustered and unclustered lineages, although this was based on a conservative estimate with only one gene missing in the *mce* operon allowed (Figure 32). This analysis also doesn't determine whether the *mce* operons are orthologous, and thus it is possible that *mce* operons are contributing to the difference in virulence of the clustered and unclustered lineages. On the other hand, other analyses have suggested that the number of *mce* operons is not associated with virulence and rather that niche specialization has led to the differing number of *mce* operons encoded by different species, the results of this analysis support this hypothesis (259, 302).

Whilst potential candidates associated with the continuing adaptation of the clustered lineages were identified, this analysis was limited by a lack of signal detected on the branches after the clonal expansion of the clustered lineages. This could be due to several factors such as: i) not enough time having passed since the clonal expansion of the lineages for selection to have occurred, ii) variants in the accessory genome could have been missed due to not mapping to a close enough reference iii) recombination could be contributing to the adaptation of the clustered lineages, iv) variants that became fixed during the early rounds of evolution, prior to the population bottleneck, are not possible to decipher from

those that occurred prior to the LCA and thus are not detected in this analysis and v) true signal could have been discounted by removing the SNPs on the terminal branches

Further research is required to quantify exactly what impact time is playing on the ability to detect whether selective pressure is acting on the branches after the clonal expansion of the clustered lineages. Bryant et al. dated the emergence of the three largest lineages, *M. a. abscessus* cluster 1 (DCC1), *M. a. abscessus* cluster 2 (DCC2) and *M. a. massiliense* cluster 1 (DCC3) to be 1980, 1963 and 1972 respectively (73). This analysis has not been performed for the remaining clusters and thus it is unknown how recently these lineages emerged, although the densely clustered nature of the clades suggests it was recent, which could mean that not enough time has passed since the emergence of these lineages for selection to have occurred. The temporal dynamics of these lineages warrant further investigation.

Another contributing factor to the lack of signal could be the loss of variants occurring in the accessory genome due to mapping the isolates firstly to their individual subspecies reference genomes and secondly to a single reference genome, *M. a. abscessus* ATCC19977. This approach was taken in order to maximize the chances of identifying genes under selection in the core genome and thus after the clonal expansion of multiple clusters. However, by mapping isolates from each cluster to a reference from within the cluster and thus increasing the resolution the power to detect genes under selection on multiple branches within a cluster may be increased. It should be noted that the *M. a. abscessus* ATCC19977 reference genome falls within *M. a. abscessus* cluster 1 and both *M. a. abscessus* candidates included nonsynonymous SNPs acquired within this cluster, suggesting that lack of resolution within the other clusters due to mapping to a genetically distant reference genome is playing a role in the lack of signal detected.

Recombination could also be playing a role in the adaptation of the clustered lineages to increased virulence and transmission. The majority of SNPs removed due to potentially being introduced via recombination were associated with mobile elements, particularly phage (Figure 33). These SNPs were also predominantly synonymous (Figure 25, Figure 27), which was to be expected as regions of the genome that have been introduced from an external source have been under a differing selection pressure where in the majority of cases deleterious changes have had time to be purged (280). However, it is not possible to definitively rule out that the SNPs removed due being acquired via recombination are not playing a role in the increased virulence and transmissibility of the clustered lineages. The

focus of this analysis was on using independently acquired SNPs to identify genes associated with the continuing adaptation of the clustered lineages and therefore the genes which overlapped the SNP dense regions were not investigated further and thus warrant further analysis in the future.

Finally, the signal was also weakened by removing the terminal branch SNPs. These were removed because the aim of this project was to identify genes that were driving the continuing expansion of the clustered lineages, and it could be argued that there is no evidence that the variants accumulated on the terminal branches have contributed to the continuing spread of the lineage as they have no descendants. Clearly this interpretation is biased by sampling and is also based on the assumption that the tree structure is correct. A way to include the terminal branch SNPs and satisfy the criteria that there is evidence that the acquisition of SNPs in this gene has led to the continued expansion of the lineage would be to investigate the genes that accumulated a significant number of nonsynonymous SNPs with the terminal branch SNPs included, but only investigate further those in which at least one variant was accumulated on the branches deeper in the clusters. This analysis should be expanded to include this in the future.

3.5 Conclusions and Future Directions

This analysis aimed to identify genes associated with the continuing expansion of the clustered lineages, which cause the majority of MABSC infections in the CF community and have been shown to be more virulent. Three candidate genes were identified with the most significant candidate being the finding that a regulator of an *mce* operon was under selection in one of the largest clusters that has spread globally. Whilst the functions of the other two candidate genes could not be determined, they provided evidence to suggest that clustered lineages from differing genetic backgrounds were adapting in the same way to enable their continued success in the CF environment. This investigation was hampered by the limited signal detected. Further analysis is required to address the effect time is playing on the ability to detect signal after the clonal expansion of the clustered lineages as well as to investigate the contribution of other types of variation such as indels, recombination and gene presence absence to the evolution of these lineages. Despite the limited signal, interesting candidates were identified and further analysis is required to determine their function, which in turn could increase our understanding of how the organisms of the MABSC are adapting to become more successful in lung environment.

3. Continuing expansion of the clustered lineages

4. Adaptation of the *Mycobacterium abscessus* species complex to the Cystic Fibrosis lung

Statement of contribution:
This project was supervised by Julian Parkhill and Andres Floto. I performed all the bioinformatics analysis reported in this chapter. RNA extraction was carried out by Daniela Rodriguez-Rincon. The phoPR knock-outs were generated by Juan Manuel Belardinelli. Julian Parkhill, Andres Floto, Mary Jackson, Josephine Bryant and Daniela Rodriguez-Rincon, Juan Manuel Belardinelli contributed to the interpretation of these results.

4.1 Introduction

People with CF have increased susceptibility to chronic pulmonary infections caused by a variety of bacterial, fungal and viral pathogens. The three subspecies of the MABSC are an emerging threat to people with CF and due to their highly antibiotic resistant nature are extremely difficult to treat resulting in treatment failure rates as high as 50% (55, 149, 306). Therefore, novel treatments are urgently needed.

MABSC organisms are found in a soil and water environment and are not natural inhabitants of the lung, although their ability to replicate in amoebae and form biofilms on surfaces suggests that adaptations beneficial to survival in their original niche have contributed to their ability to cause opportunistic infections (92, 93, 95, 102-104). However, how the MABSC has specifically evolved in response to the selection pressures applied by the host within the CF lung environment is not well understood and a greater understanding of this would increase our knowledge of the pathogenesis of the MABSC, which could in turn potentially uncover novel drug targets.

Examining the evolution of bacteria over time within its host has long been used to determine the genes and pathways involved in the adaptation of an opportunistic pathogen to a novel niche. This approach has been used extensively to examine the within patient evolution of other CF pathogens, including *Pseudomonas aeruginosa* (182, 307-309), *Staphylococcus aureus* (309, 310) and the *Burkholderia cepacia* species complex (40, 311, 312). In these analyses, parallel adaptive evolution was observed in antibiotic resistance associated genes as well as genes involved in bacterial membrane composition, metabolism, biofilm formation and regulation (308, 310, 312). Furthermore, possible novel treatments were identified, for example the possibility of targeting the heme utilisation pathway of *P. aeruginosa* (307).

Thus far applying longitudinal analysis to the MABSC has been limited by small sample sizes and a lack of longitudinal samples and has predominantly been undertaken with the aim of identifying antibiotic resistance mutations (313). Although a couple of studies have begun to investigate the adaptation of the MABSC to the CF lung through this method: Kreutzfeldt and colleagues sequenced 178 isolates obtained from 12 patients, over a period of 12 years. However, they only examined the genetic changes over time with regards to adaption to the lung in one patient. Through this they identified nonsynonymous mutations in genes involved in transcriptional regulation and metabolism and most interestingly the acquisition of two independent nonsynonymous mutations in the histidine kinase encoding *phoR* gene, which is

part of the PhoPR two component system (TCS) (183). The PhoPR TCS regulates a myriad of genes involved in complex lipid biosynthesis and virulence in *M. tuberculosis* and is believed to be critical to its pathogenicity (314, 315). Davidson and colleagues in their analysis of the MABSC global population structure also sequenced multiple isolates per patient and examined the within host evolution, however, their analysis of the longitudinal samples only extended as far as to comment on the level of within host diversity and did not extend to identifying parallel evolution between patients in order to identify genes associated with the adaptation to the lung (89).

Consequently, given that the detection of parallel evolution in multiple patients has been shown to enable the discovery of genes and pathways associated with the adaptation of saprophytic bacteria to a human host and that this approach has yet to be applied to the MABSC, the aim of this project was to detect parallel evolution between patients infected with MABSC infections in order to provide insights that could potentially increase our knowledge of the pathogenesis of the MABSC, identify novel drug targets and inform intervention strategies.

4.2 Materials and Methods

4.2.1 Within host evolution dataset

For 201 patients from which samples were collected for MABSC global collection more than one isolate was obtained. These isolates collected from multiple patients could either represent co-infection with genetically diverse lineages, a transmitted lineage or the within host evolution of a single lineage. In order to determine the isolates in which the genetic changes would have occurred under selection pressure from the host, phylogenetic analysis was performed.

4.2.2 Mapping, variant calling and de novo assembly

All 1252 isolates in the MABSC global population dataset were mapped to *M. a. abscessus* ATCC19977 using BWA-MEM (v. 0.7.2), using the parameters described in section 7.3 (215). Variants were called and extracted from the alignments following the methods described in section 7.3 and 7.4. *De novo* assembly and annotation for each isolates was carried out as described in section 7.6 and 7.7.

4.2.3 Phylogenetic analysis

To ensure that the variants being counted were only those that had occurred under within host selection pressure, it was necessary to determine and remove isolates from patients for which the last common ancestor was that of another patients and not another isolate from the same patient. To achieve this a maximum likelihood phylogenetic tree was inferred, using RAxML (v.8.2.8), from an alignment of the variant cites. From this phylogeny, only isolates from a patient that either formed a monophyletic clade (Appendix Figure 3.1) or a paraphyletic clade in which another patients isolates were nested within the diversity of the patient of interest (Appendix Figure 3.1) were included in the longitudinal dataset. This resulted in a final longitudinal dataset of 810 isolates obtained from 182 patients (Appendix Table 3.1).

4.2.4 Detecting genes that accumulated a significant number of nonsynonymous SNPs in multiple patients

To determine whether a gene had accumulated a greater than expected number of nonsynonymous or nonsense SNPs, the 'burden of mutation' method described by Ding et al. (2008) was used as described in section 7.11 (295). For this method to work recombination has to be removed, this was achieved by identifying blocks of three or more SNPs that occurred within 1000bps of each other and then identifying if the isolates from the patient either accumulated all the SNPs in the possible recombination or did not accumulate any of the SNPS in the possible recombination. If this criteria was met, these SNPs were removed as potential recombination.

4.2.5 Candidate gene follow up analysis

The Prokka annotations of the candidate genes were enhanced by searches against the Pfam (v.3.1.0) and InterPro (v.68) protein databases (224, 225). Candidates possibly associated with phage were identified using the PHASTER database (228). Candidate genes orthologous to *M. tuberculosis* H37Rv were identified from the catalogue of Mycobacterium orthologs described in (230). The functional interpretation of the candidates was also enhanced by pathway and functional enrichment analysis using the STRING database and the webtool DAVID (6.8) .

For more specific analysis of the candidates, the protein sequences of some of the candidates were aligned to their *M. tuberculosis* H37Rv ortholog using muscle (v.3.8.31) (60). To determine if a specific domain of a gene was under selection, where enough nonsynonymous SNPs had been accumulated for this analysis to be possible, a permutation

test approach was used to generate the expected distribution of nonsynonymous SNPs amongst the domains, followed by a Fisher's exact test (two tailed) to determine if there was a significant difference between the observed and expected distribution of nonsynonymous SNPs amongst the domains. A p-value of less than 0.01 was seen as significant.

4.2.6 Generation of PhoPR knockout mutants

The following work was performed by Juan Manuel Belardinelli. Homologous recombination at the PhoPR locus of *M. a massiliense* CIP108297 was performed using a mycobacterial recombinase-based system in which the recombineering genes from mycobacteriophage Che9c are expressed from the replicative plasmid pJV53-xylE (a derivative of the pJV53 plasmid generated in-house in which the xylE colored marker was added to improve selection of transformants) under control of an acetamide-inducible promoter (316). Acetamide-induced *M. a. massiliense* CIP108297 cells harboring pJV53-xylE were electro-transformed with approximately 300 ng of linear allelic exchange substrate consisting of the streptomycin-resistance cassette from pHP45Ω flanked by 1,000 bp of DNA sequence immediately flanking the PhoPR operon, and double-crossover mutants were isolated on Str-containing agar. Allelic replacement leading to the complete deletion of the PhoPR locus was checked by PCR using a pair of primers annealing outside the linear allelic exchange substrate.

4.2.7 RNA sequencing and differential expression analysis

RNA sequencing was carried out on the illumina V4 platform to generate 75bp paired end reads. The reads were mapped to *M. a. abscessus* ATCC19977 using BWA v.0.7.112 to produce a bam file, with BWA also used to index the reference and align the reads (317). The default parameters were used apart from the quality threshold for read trimming which was set to 15 and the maximum insert size which was set at 510bp, the maximum fragment size of the sequencing library. Duplicate reads were marked using Picard (318).

Gene expression values were computed from alignment of the reads to the CDSs present in the *M. a. abscessus* ATCC19977 chromosome. These were used to generate the reads mapping and reads per kilobase per million (RPKM) . Only reads with a mapping quality score of 10 were included in the count (319, 320). In order to follow a strand-specific approach, a read which was a true pair and which was the second of a pair was modified depending on which strand it mapped to. If the read met the aforementioned criteria and mapped to the forward strand the read was modified to map to the reverse strand, with the read sequence not being complemented, whilst if a read mapped to the reverse strand then

the read was modified to map to the forward strand with the read sequence not complemented.

Differential expression analysis was carried out using the R package DESeq2 (321). Briefly, DESeq2 takes the raw count data, broken down per gene, from an RNAseq experiment and proceeds to correct the counts for library size and dispersion, identifying significantly differentially expressed genes using the Wald statistical test, which tests the null hypothesis that the log fold change between the gene under the two conditions is negligible (321). The p-values were corrected for multiple testing using the Benjamini Hochberg method. P-values of less than 0.01 were seen as significant. The analysis were carried out using the Deago pipeline (322).

4.3 Results

The final longitudinal dataset consisted of 810 isolates obtained from 182 patients (Table 11). All three subspecies were represented, with 496 *M. a. abscessus* isolates sampled from 123 patients, 86 *M. a. bolletii* isolates sampled from 14 patients and 228 *M. a. massiliense* isolates sampled from 46 patients. On average 4 isolates were sampled per patient. For one patient, SMRL_J, multiple isolates from two separate subspecies, *M. a. massiliense* and *M. a. abscessus*, were sampled. For three other patients, PAP_007, SMRL_AT, and PAP_019, the evolution of two separate *M. a. massiliense* lineages were examined. Overall the within host evolution was examined for 186 MABSC lineages.

Table 11: Summary of the final within host evolution dataset

Total number of isolates:		810
Total number of patients:		182
Total number of lineages following evolution of:		186
Patients with multiple lineages:		SMRL_J (1xmass 1xabss)
		PAP_019 (2xmass)
		SMRL_AT (2xmass)
		PAP_007 (2xmass)
M. a. abscessus	**patients**	123
	isolates	496
M. a. bolletii	**patients**	14
	isolates	86
M. a. massiliense	**patients**	46
	isolates	228

A total 1,185 SNPs were accumulated over time within 137 patients, with no SNPs accumulated by lineages within 45 patients (Appendix table 3.2). Overall 682 nonsynonymous or nonsense SNPs, 369 synonymous SNPs and 134 intergenic SNPs were detected (Appendix table 3.4). An average of 6.3 SNPs were accumulated per lineage per patient. To see if the lineages within patients were evolving in a clock like manner, a linear

regression plot was constructed for the 116 lineages for which the day/date/year of sampling was available (Appendix table 3.3). Figure 34A showed that there was a weak positive correlation ($r^2 = 0.24$) between the number of SNPs accumulated and the time span of collection. It should be noted that this does not take into account time since infection but before sampling. The slope coefficient, 0.010211, represents the number of substitutions accumulated per day which equates to the accumulation of 3.7 substitutions per year. However, there was a clear outlier that accumulated significantly more SNPs than would have been expected and which couldn't be explained by sampling time or recombination. Given that the acquisition of a hypermutator phenotype is a feature that has been associated with other CF pathogens and has been observed before for MABSC lineages (Josephine Bryant, unpublished data), the isolates associated with this patient were investigated for possible SNPs that could infer a hypermutator phenotype.

4.3.1 Acquisition of a hypermutator phenotype

One of the subclones from patient SMRL_AT stood out as having accumulated significantly more SNPs over time, having accumulated 135 SNPs over a time span 3 years, 11 months and 26 days, a 36 fold increase in the substitution rate (Figure 34A). No nonsynonymous or nonsense SNPs were identified in genes that could potentially confer a hypermutator phenotype and therefore indels were investigated to look for potential frameshift mutations. The deletion of the adenine at base position 77 in the endonuclease III (nth) gene (MAB_0418), encoding a DNA repair enzyme that excises damaged pyrimidines from double stranded DNA and which has been shown to confer a hypermutator phenotype in *E. coli*, was identified (323, 324). Removing this patient from the linear regression analysis resulted in a slight increase in the correlation coefficient (adj R^2 0.43) and a slope coefficient of 0.009 (Figure 34B), resulting in the estimated substitution rate being reduced to the accumulation of 3.1 SNPs per year, which overlaps with previously estimated substitution rates for this species (70).

A hypermutator phenotype introduces many variants into the population, some of which may be beneficial to the organism's survival. Therefore, in order to increase the signal and despite the potential for hitchhiker mutations, the lineage from the patient with the hypermutator phenotype was investigated, along with the lineages from the remaining patients, for evidence of parallel evolution occurring between patients.

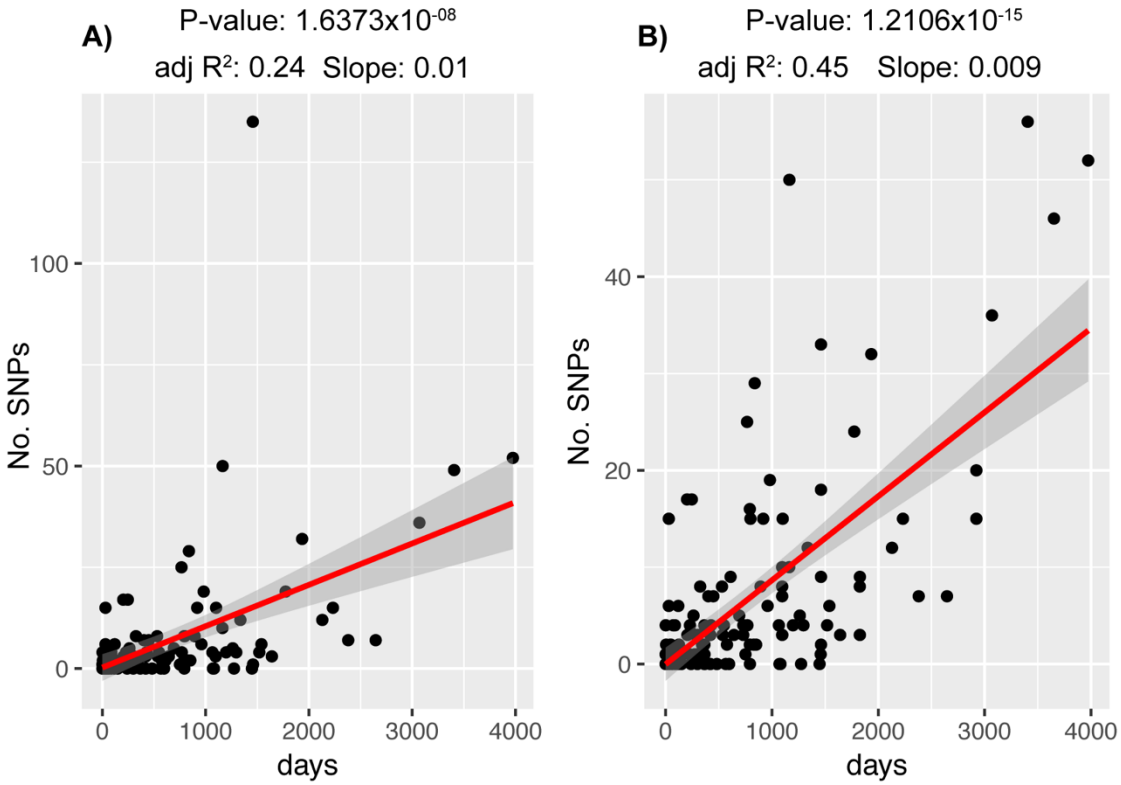


Figure 34: MABSC can acquire a hypermutator phenotype over time within patients

Linear regression analysis showing the number of variants accumulated within a patient plotted against the time span between the earliest and latest sample available for the patient (n=116). A) shows that whilst the majority of patients are accumulating SNPs in a clock-like manner, one patient (SMRL_AT) has accumulated significantly more SNPs that would have been expected given the time span over which the samples were collected. B) shows the affect that the removal of the patient that has acquired a hypermutator phenotype has on the correlation. The grey shaded represents the 95% confidence interval for the regression.

4.3.2 MABSC lineages evolving in parallel within the CF lung in multiple patients

A nonsynonymous or nonsense mutation was acquired by 125 of the 186 (67%) lineages (Appendix table 3.2). The SNPs were distributed across 461 genes (including 16s rRNA and 23s rRNA), with 61 genes accumulating nonsynonymous or nonsense SNPs in more than one patient over time (Appendix table 3.4). Seventeen of the 61 genes were found to have accumulated a greater number of nonsynonymous or nonsense SNPs in multiple patients than would have been expected by chance (adjusted p-value < 0.01) (Table 12), suggesting that these genes were potentially involved in the adaptation of the MABSC to the CF lung (full results can be seen in Appendix table 3.5). In total 128 nonsynonymous or nonsense mutations was accumulated by the 17 candidate genes, with SNPs observed in these genes in 58 of the 182 patients (32%) investigated.

Table 12: Summary of the genes which acquired a significant number of nonsynonymous SNPs in multiple patients

Locus	Product		*M. tuberculosis* locus	Product		No. Observed nonsyn SNPs	Expected no. nonsyn SNPs	No. of patients	Binomial p-value	Adj. p-vlaue
MAB_0173	Prokka	Prenyltransferase family protein	Rv3806c	UbiA		11	0.21	11	0	0
	Pfam	UbiA								
	InterPro	UbiA								
MAB_2372	Prokka	Probable GTP-binding protein EngA	Rv1713	EngA		12	0.314	10	0	0
	Pfam									
	InterPro									
MAB_3539	Prokka	Putative transcriptional regulator, WhiB family	Rv3219	WhiB1		7	0.064	7	6.22E-15	1.02E-11
	Pfam									
	InterPro									
MAB_1881c	Prokka	Putative transcriptional regulator, TetR family	NA	NA		8	0.157	7	1.35E-13	1.66E-10
	Pfam									
	InterPro									
MAB_0674	Prokka	Putative sensor histidine kinase PhoR	Rv0758	PhoR		20	2.693	15	1.28E-12	1.26E-09
	Pfam									
	InterPro									
MAB_0416c	Prokka	Putative CRP transcriptional regulator	Rv3676	cAMP receptor binding protein		7	0.148	7	4.90E-12	4.02E-09
	Pfam									
	InterPro									
MAB_3029	Prokka	Iron-dependent repressor IdeR	Rv2711	IdeR		7	0.159	4	8.64E-12	6.07E-09
	Pfam									
	InterPro									
MAB_1068c	Prokka	Conserved hypothetical protein	Rv0966c	Tuberculist	Conserved hypothetical	6	0.141	5	1.87E-10	1.15E-07
	Pfam	DUF1707 – probably membrane associated		Pfam	DUF1707					
	InterPro	DUF1707 – probably membrane associated		InterPro	DUF1707					
MAB_0115c	Prokka	Conserved hypothetical protein	Rv3849	EspR		5	0.096	5	9.57E-10	5.23E-07
	Pfam	No domains predicted								
	InterPro	One unintegrated signature								
MAB_0189c	Prokka	Probable arabinosyltransferase C	Rv3793	EmbC		15	2.806	13	4.50-08	2.21E-05
	Pfam	Mycobacterial cell wall arabinan syntehesis protein/ EmbC C-terminal domain								
	InterPro	Arabinofuranosyltransferase, N-terminal domain, Arabinosyltransferase, C-terminal domain								

MAB_3817c	Prokka Pfam InterPro	50S ribosomal protein L2	Rv0704	50S ribosomal protein L2	5	0.189	5	5.22E-08	2.33E-05
MAB_3036c	Prokka Pfam InterPro	Conserved hypothetical protein ATP cone domain – regulatory domain ATP cone domain, ribonucleotide reductase regulator (NrdR-like)	Rv2718c	NrdR	4	0.106	3	1.01E-07	4.13E-05
MAB_4695c	Prokka Pfam InterPro	Putative glycosyltransferase/rhamnosyltransferase No domains predicted Glycosyltransferase signatures	Rv1524/Rv1526c	NA	5	0.277	5	4.88E-07	0.0002
MAB_1483	Prokka Pfam InterPro	Conserved hypothetical protein Beta-lacatamase superfamily domain Beta-lacatamase superfamily domain	Rv1339	tuberculist: Conserved hypothetical protein Pfam: MBL-fold metallo hydrolase domain InterPro: Ribonuclease Z/Metallo-beta-lactamase	4	0.165	4	8.81E-07	0.0003
MAB_1915	Prokka Pfam InterPro	Probable fatty acid-CoA ligase FadD	NA	NA	5	0.408	4	4.46E-06	0.001
MAB_2297	Prokka Pfam InterPro	Probable methyltransferase (*erm* (41)) (*erm* (41))	NA	NA	3	0.104	3	4.42E-06	0.001
MAB_4532c	Prokka Pfam InterPro	Conserved hypothetical protein Acetyltransferase (GNAT) domain/sterol carrier protein domain N-acetyltransferase Eis	Rv2416c	Eis2	4	0.282	2	1.15E-05	0.003

Fourteen of the 17 genes were predicted to be orthologous to genes in the *M. tuberculosis* H37Rv genome (Table 12). Functional overlap between the candidate genes was investigated by querying the 17 genes against STRING database, which showed no pathways were enriched. The webtool DAVID clustered the seven genes with regulatory functions, however, the clustering was not significant (medium stringency, Score: 1.34; all p-values > 0.05). Despite the lack of significant enrichment of genes in particular pathways or functional categories, genes with similar functions and involved in overlapping pathogenic processes were evident (Figure 35).

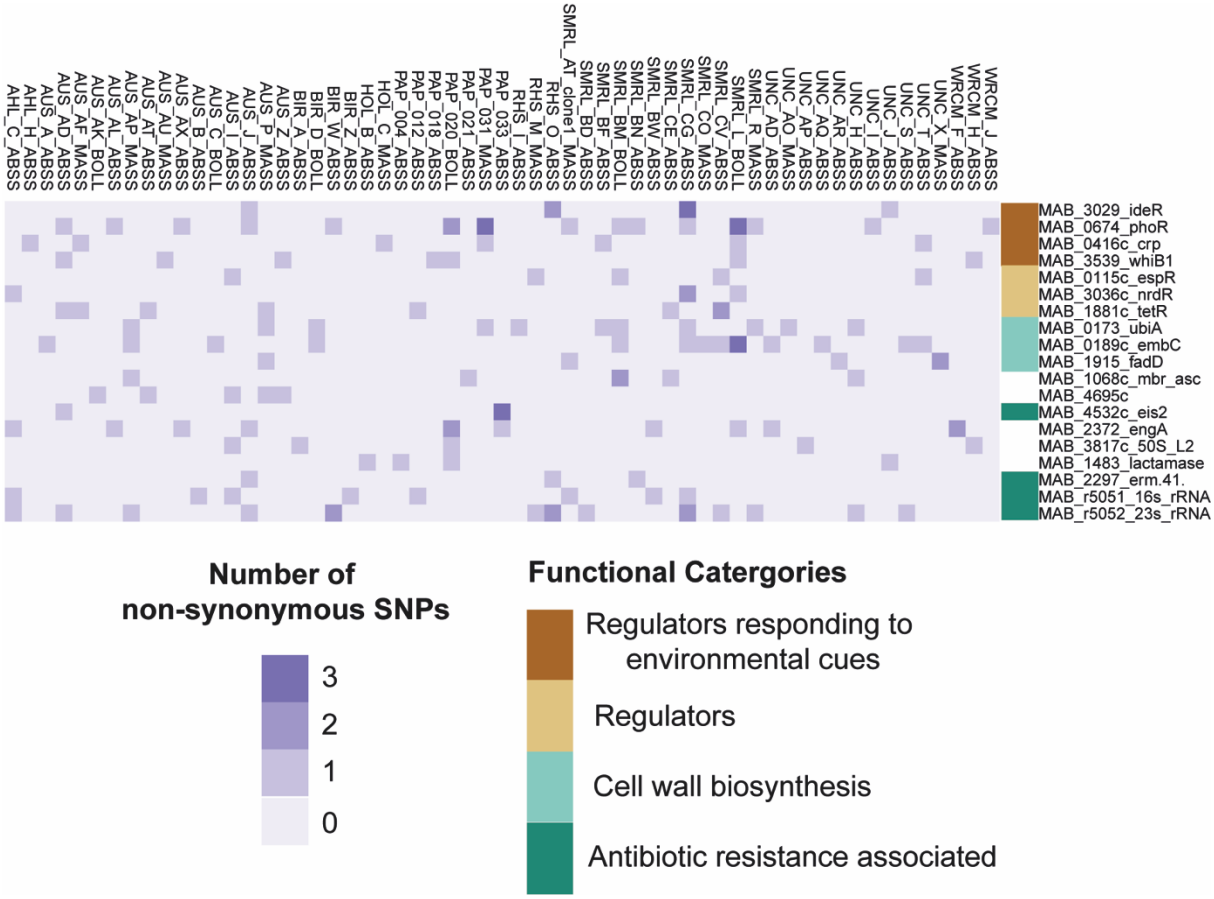


Figure 35: Functional similarity between genes evolving in parallel in multiple patients

Heat map showing the number of nonsynonymous and nonsense SNPs accumulated per patient in the 17 genes that accumulated a greater number of nonsynonymous SNPs than would have been expected by chance. The variants accumulated over time in 16s rRNA gene (MAB_r5051) and 23s rRNA gene (MAB_r5052), that are known to be associated with aminoglycoside and macrolide antibiotic resistance respectively are also shown. Out of the 182 patients for which the within patient evolution of an MABSC lineage was investigated 59 (32%) accumulated at least one nonsynonymous or nonsense mutation in either one of the 17 candidate genes or the 16s rRNA or 23s rRNA genes. The candidate genes are grouped according to possible overlapping functions: regulators stimulated by environmental cues (brown), other regulators (beige), genes involved in cell wall biosynthesis (light green) and antibiotic resistance associated (dark green).

4.3.3 Selection for variants in regulators which respond to environmental cues

Four genes which directly or indirectly sense environmental cues in the phagosome and respond by causing subsequent large scale changes in gene expression accumulated a significant number of nonsynonymous SNPs over time within multiple patients (Figure 35). These were MAB_3029, an iron dependent repressor (IdeR), MAB_0674, the histidine sensor kinase component, PhoR, of the PhoPR TCS; MAB_0416c, a cyclic AMP receptor binding protein (CRP) and MAB_3539, a WhiB family regulator. All four genes were found to be orthologous to genes in *M. tuberculosis* H37Rv, with *ideR* (MAB_3029) being orthologous to Rv2711, *phoR* (MAB_0674) being orthologous to Rv0758, CRP (MAB_0416c) being orthologous to Rv3676 and MAB_3539 (from herein referred to as *whiB1*) being orthologous to the WhiB1 regulator Rv3219 (Table 12) (230).

Twenty-eight different patients accumulated at least one nonsynonymous SNP in one of these regulators sensing phagosomal environmental cues, with six patients accumulating nonsynonymous SNPs in more than one of these genes (Figure 35). Amongst these four genes, the most nonsynonymous SNPs, 20, were accumulated by the *phoR* component of the PhoPR TCS, whilst CRP, *whiB1* and *ideR* accumulated seven each. Fifteen patients accumulated one or more nonsynonymous mutation in *phoR*, whilst seven patients accumulated nonsynonymous SNPs in CRP and *whiB1* and 5 patients accumulated nonsynonymous SNPs in *ideR*. In order to hypothesize what impact the acquisition of these nonsynonymous SNPs could be having of the function of these genes, the distribution of the nonsynonymous SNPS across each gene was investigated to see if particular domains were under selection.

4.3.4 Within host selection pressure acting upon specific domains within the regulators responding to environmental cues

The protein structures of IdeR, PhoR, CRP and WhiB1 in *M. a. abscessus* ATCC19977 are unknown, however, the structures of their orthologs in *M. tuberculosis* H37Rv have been determined. The amino acid alignments between these genes in *M. a. abscessus* ATCC19977 and *M. tuberculosis* H37Rv showed a high level of AAI between three of the genes with their *M. tuberculosis* H37Rv ortholog, with an AAI of 82.5% between IdeR protein sequences, MAB_3029 and Rv2711, 96% between CRP protein sequences, MAB_0416c and Rv3676 and 89% between WhiB1 regulators MAB_3539 and Rv3219. Contrastingly, the AAI between the PhoR encoding MAB_0674 and Rv0758 was 32%. However, the high level sequence identity between three of the four ortholog pairs and the presence of identical conserved motifs amongst all of them, suggested that it would be possible to potentially use

the functional information known for these genes in *M. tuberculosis* H37Rv to predict the effect of the nonsynonymous mutations in their corresponding orthologs in *M .a. abscessus* ATCC19977.

The structure of IdeR in *M. tuberculosis* H37Rv has shown that it forms a homodimeric protein, with each monomer consisting of a helix-turn-helix (HTH) diphtheria toxin regulator (HTH_DTXR) domain, predicted to be encoded by amino acids 26 to 125 in MAB_3029 and an FeoA (dimerization) domain, which was predicted to be encoded by amino acids 151 to 227 in MAB_3029 (Figure 36). A third domain similar to the Src-SH3 domain is also encoded (325). Three of the seven nonsynonymous mutations accumulated in parallel in this gene were observed in HTH_DTXR domain, with one falling within the HTH DNA binding domain and one in the FeoA domain, however, none of the nonsynonymous SNPs occurred at the metal ion binding sites (site 1: His^{79}, Glu^{83}, His^{98}, $Glu1^{72}$, Gln^{175}; site 2: Met^{10}, Cys^{102}, Glu^{105} and His^{106}) or at sites predicted to make contact with DNA (325, 326).

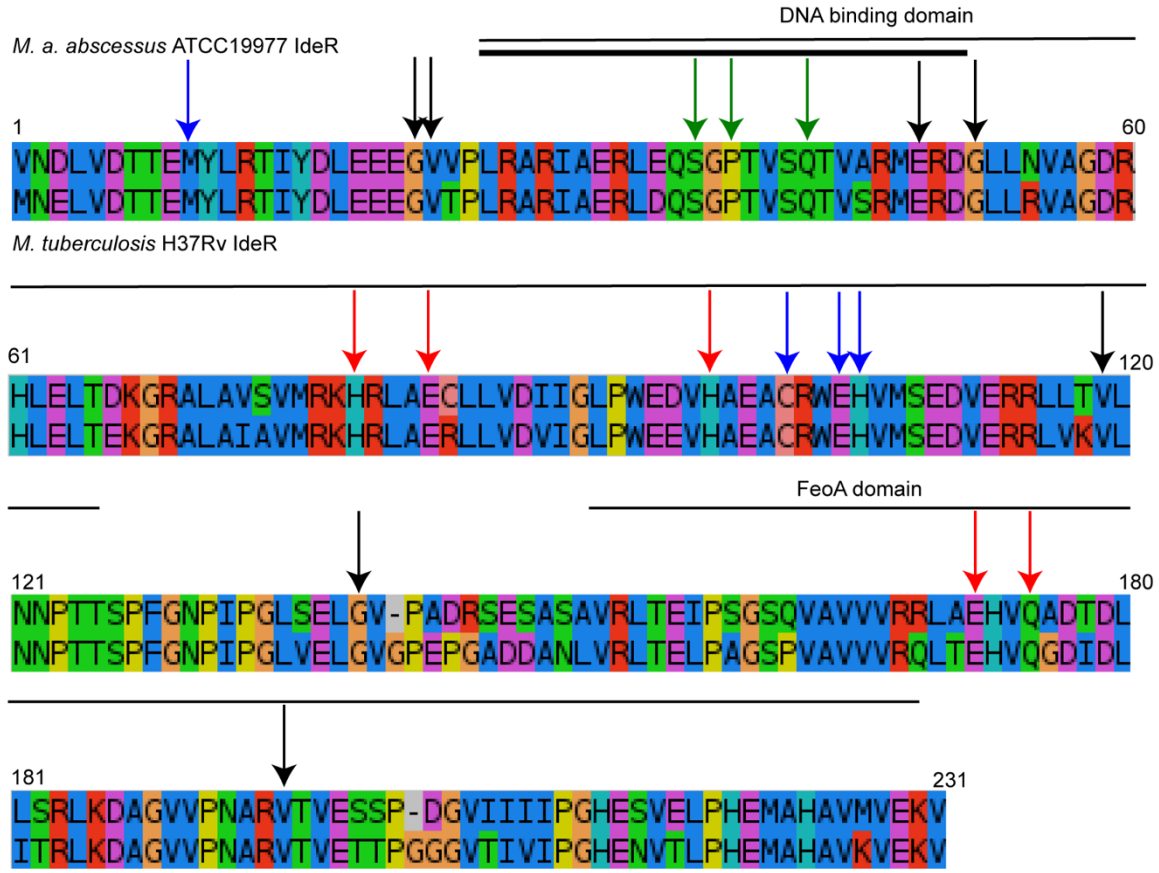


Figure 36: Nonsynonymous SNPs acquired in parallel in IdeR functional domains

Alignment of the IdeR protein sequences from *M. a. abscessus* ATCC199777 and *M. tuberculosis* H37Rv, adapted from (327). **Black** arrows represents the positions of the nonsynonymous SNPs accumulated in parallel between patients infected with MABSC lineages. Red arrows indicate amino acids that interact with the metal ion in metal ion binding site 1. Blue arrows indicate amino acids that interact with the metal ion in metal ion binding site 2. Green arrows indicate sites predicted to make contact with DNA. The thin black lines represent the DNA binding and FEO domains respectively as predicted by SMRT. The bold black line marks the helix turn helix DNA binding motif (aa 25-51).

CRP, which consists of a cAMP binding domain, predicted to be encoded by amino acids 10 to 128 in MAB_0416c, and a helix-turn-helix cAMP regulatory domain, predicted to be encoded by amino acids 167 to 215, acquired 7 nonsynonymous SNPs in parallel (Figure 37). Three of the SNPs, E80G and R89Q in two patients, were acquired in the cAMP binding domain at sites known to directly bind cAMP in *M. tuberculosis* H37Rv (328). Furthermore, the mutation observed in the HTH cAMP regulatory domain, G185S, occurred at a highly conserved site (328).

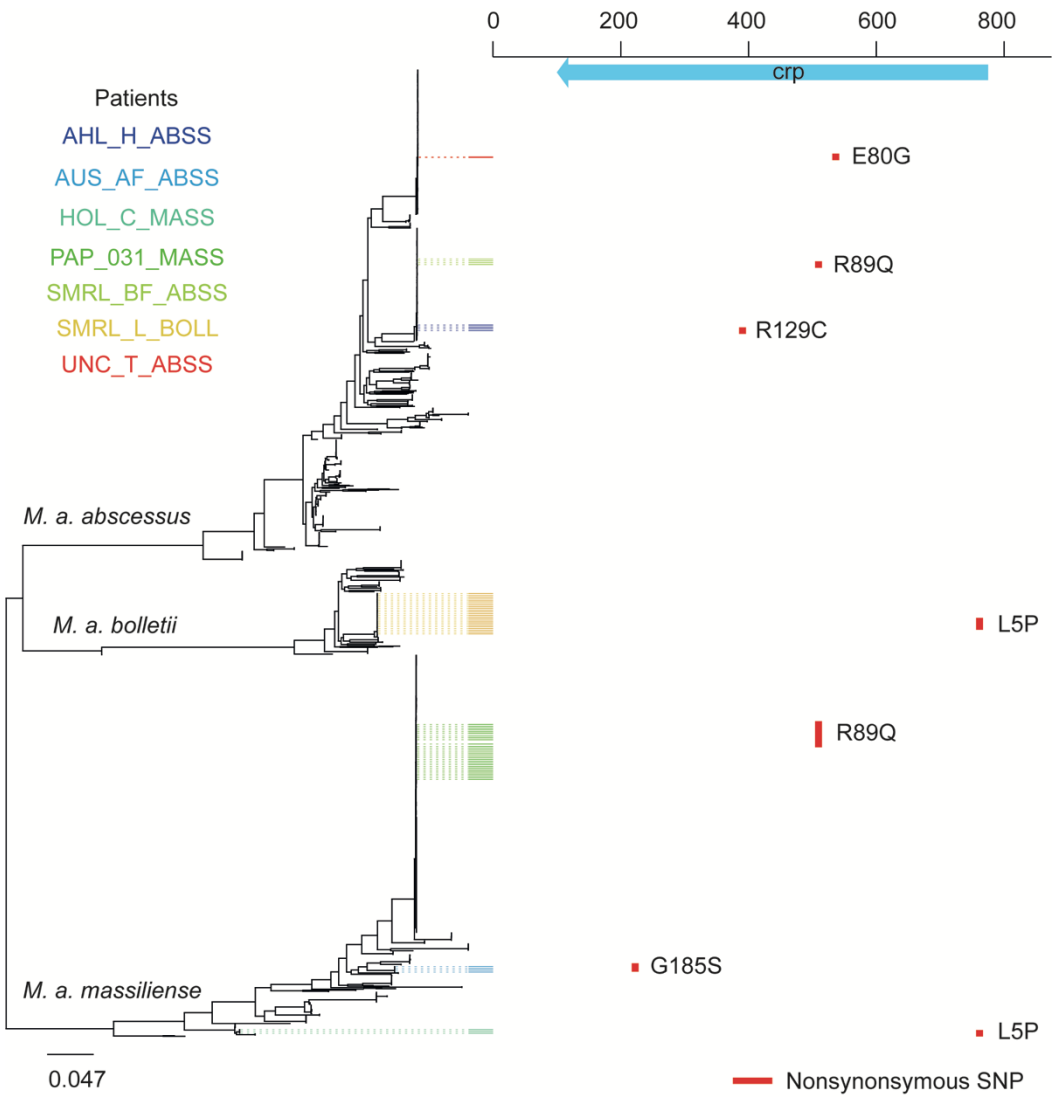


Figure 37: Variants were acquired in parallel in CRP at sites that could potentially affect its function.

The MABSC global population structure showing that variants were accumulated by CRP in seven patient infected with genetically distinct lineages. Three of the variants acquired by CRP in parallel occurred at sites, R89Q and E80G, known to interact with cAMP in *M. tuberculosis*. One variant, G185S, occurred at a highly conserved site in the cAMP regulatory domain.

WhiB1 consists of a single WhiB domain, predicted to be encoded by amino acids 3 to 71 in MAB_3539. All seven nonsynonymous SNPs fell within this domain, although MAB_3539 is only 84 amino acids in length. One of the nonsynonymous mutations, R68C, occurred within a predicted helical motif, encoded by amino acids 65 to 76, whilst one of the nonsynonymous mutations, W49R, occurred at a site that forms part of the mouth of a channel where a cluster sulfide atom is exposed (329). The disruption of the structure of the channel by the nonsynonymous SNP at this site, could potentially enable NO to access the [4Fe-4S] cluster causing it to be destabilized (Figure 37) (329, 330). The destabilization of the [4Fe-4S]

cluster causes WhiB1 to disassociate from the major sigma factor, enabling the positively charged residues in the c-terminal helix, none of which accumulate a nonsynonymous change in parallel, to bind DNA resulting in a change in gene expression (329).

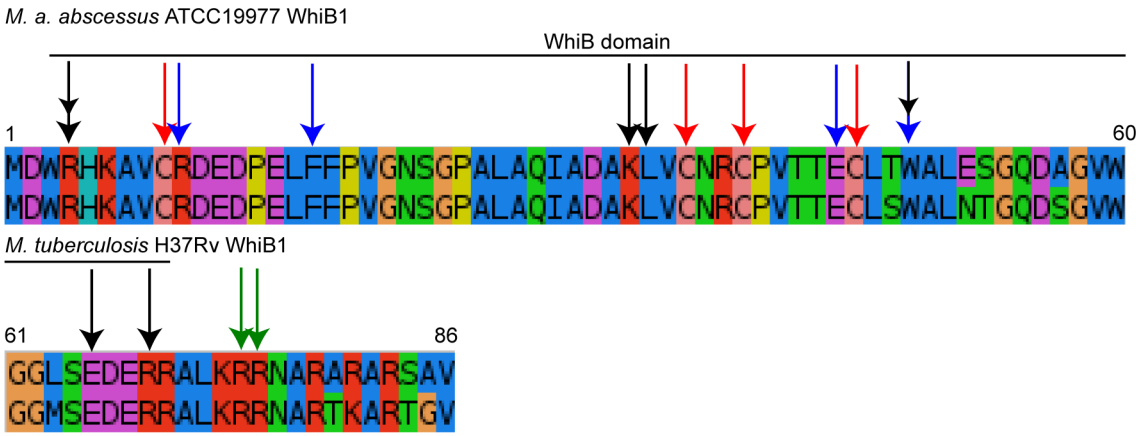


Figure 38: WhiB1 accumulated mutations in parallel at sites that could cause changes it its function

WhiB1 protein sequences of *M. a. abscessus* ATCC19977 and *M. tuberculosis* H37Rv aligned with muscle (60). Black arrows represent nonsynonymous SNPs accumulated over time within multiple patients. Red arrows indicate cysteine residues which coordinate the [4Fe-4S] cluster. Green arrows indicate the two arginine residues that are involved in DNA binding. Blue arrows indicate the residues which form the mouth of the channel through which the cluster sulfide atom is exposed (329).

Twenty nonsynonymous SNPs were accumulated in parallel by PhoR, the response regulator of the PhoPR TCS. PhoR is membrane bound protein with a sensor loop, encoded by amino acids 59 to 176 in MAB_0674, that extends into the periplasmic space, and multiple cytoplasmic domains (314). Figure 39 shows that the 20 nonsynonymous SNPs accumulated in parallel by PhoR did not appear to be randomly distributed across the protein. By mapping the nonsynonymous SNPs onto the secondary structure (Figure 40), 70% (14/20) of the nonsynonymous SNPs were found to have been accumulated in the region predicted to encode the sensor loop. Through a permutation test this was shown to be a pattern that was not expected by chance (p-value: 0.0036).

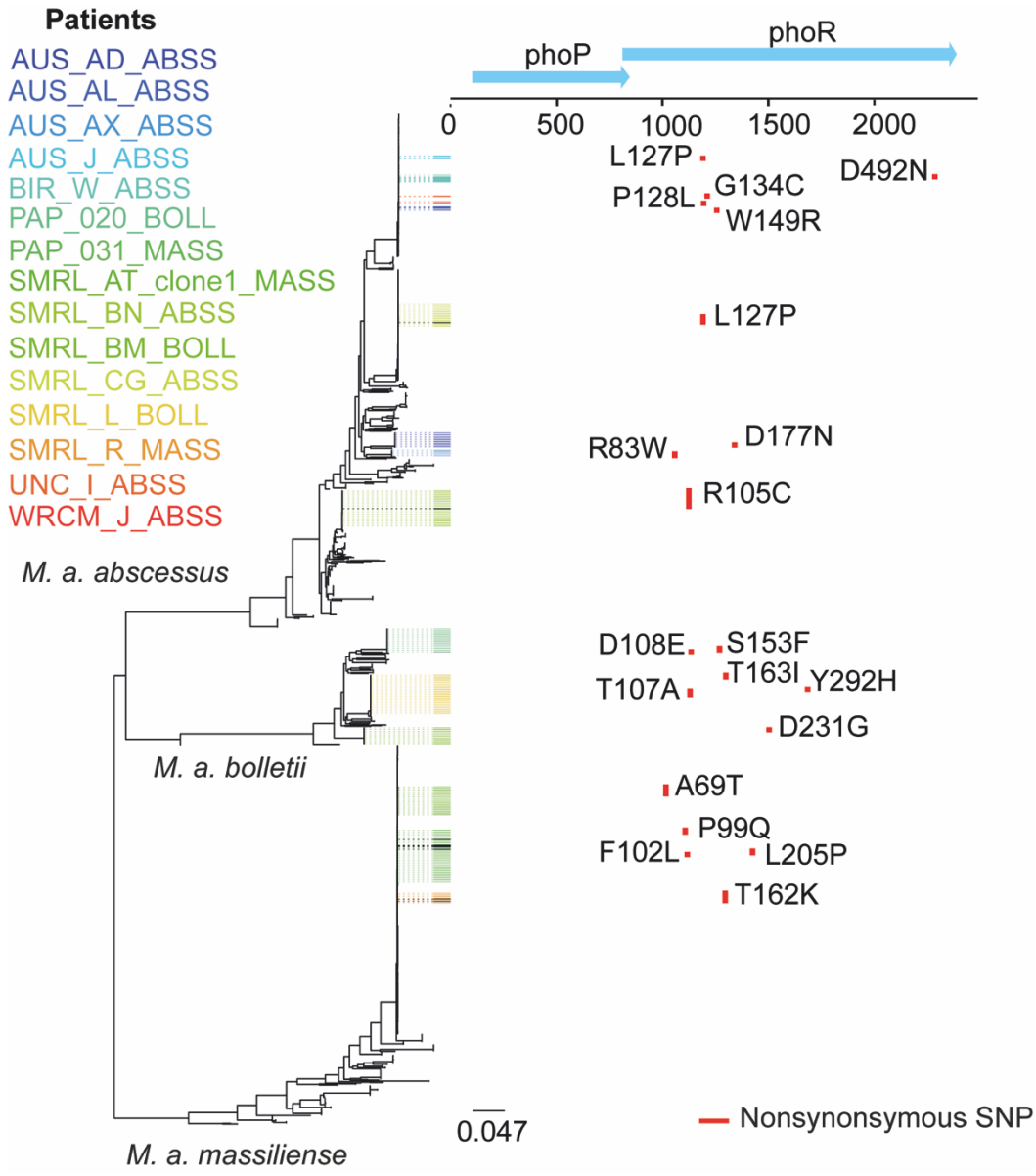


Figure 39: The 20 nonsynonymous SNPs acquired by PhoR were not randomly distributed across the protein sequence

This figure shows the MABSC global population structure with the aligned metadata representing the distribution of the 20 nonsynonymous SNPs accumulated in parallel by PhoR, the response regulator component of the PhoPR TCS, in lineages infecting 14 patients. Isolates from each subspecies accumulated nonsynonymous mutations over time within the host. The pattern of the mutations across the gene also suggested that the mutations were not randomly distributed.

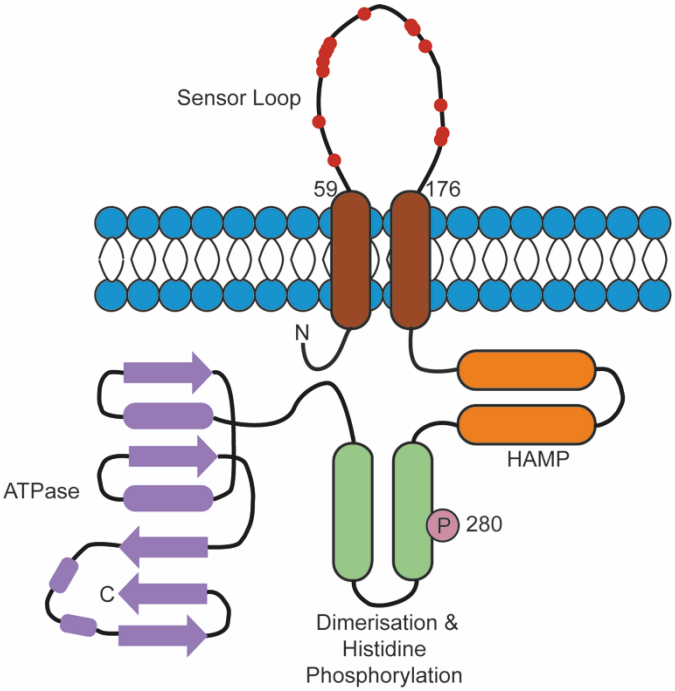


Figure 40: PhoR sensor loop under selection within the host

The secondary structure of PhoR, a histidine sensor kinase, showing its membrane topology, Adapted from Broset et al. 2015 (314). Each domain is colored separately, oblong shapes represent alpha helices, arrows represent Beta pleated sheets. 14 of the 20 (70%) nonsynonymous SNPs that were accumulated in parallel in 20 patients fell within the sensor loop, a pattern that would not be expected by chance (p-value: 0.0036).

Whilst there was evidence that within host selection pressure was specifically selecting for changes in particular functional domains in all the regulators responding to environmental cues, this was most evident for PhoR . Therefore, to understand more about the functional role that PhoR was playing in the adaptation of the MABSC to the CF lung and with the eventual aim to understand the functional impact of the changes accumulated in the sensor loop, RNA-seq analysis was performed to determine the MABSC PhoPR regulon.

4.3.5 PhoPR potentially up-regulates virulence related genes in response to a low environmental pH and carbon source of pyruvate

The role of PhoPR in *M. tuberculosis* H37Rv has been extensively studied and it has been demonstrated that the PhoPR TCS acts a key regulatory switch by sensing the point at which during the maturation of the phagosome the pH drops and pyruvate becomes readily available as a carbon source (331). As the environmental cues that PhoR is responding to in the MABSC are unknown, it was hypothesized that it could potentially be sensing similar environmental cues as PhoR in *M. tuberculosis* H37Rv. To test this hypothesis the PhoPR

TCS was knocked out in *M. a. massiliense* CIP108297 and the gene expression levels were subsequently compared between isolates in which PhoR was hypothesized to be stimulated (a combination of a low pH (5.7) and sole carbon source of pyruvate) and those in which they were not. A summary of the comparisons are in Table 13.

Table 13: Summary of the differential expression analyses performed to determine the MABSC PhoPR regulon

Comparison			phoPR stimulation	No. of activated genes	No. of inactivated genes
PhoPR WT vs PhoPR KO: CS: Glycerol pH: 5.7			not stimulated vs not stimulated	3	41
PhoPR WT vs PhoPR KO: CS: Glycerol pH: 7			not stimulated vs not stimulated	2	42
PhoPR WT vs PhoPR KO: CS: Pyruvate pH: 5.7			stimulated vs not stimulated	45	27
PhoPR WT vs PhoPR KO: CS: Pyruvate pH: 7			not stimulated vs not stimulated	3	42
PhoPR WT: CS: Pyruvate pH: 5.7	vs	PhoPR WT: CS: Pyruvate pH: 7	stimulated vs not stimulated	12	8
PhoPR WT: CS: Glycerol pH: 5.7	vs	PhoPR WT CS: Glycerol pH: 7	not stimulated vs not stimulated	0	21
PhoPR WT: CS: Pyruvate pH: 5.7	vs	PhoPR WT: CS: Glycerol pH: 5.7	stimulated vs not stimulated	16	0
PhoPR WT: CS: Pyruvate pH: 5.7	vs	PhoPR WT: CS: Glycerol pH: 7	stimulated vs not stimulated	18	23
PhoPR WT: CS: Pyruvate pH: 7	vs	PhoPR WT CS: Glycerol pH: 5.7	not stimulated vs not stimulated	13	4
PhoPR WT: CS: Pyruvate pH:7	vs	PhoPR WT: CS: Glycerol pH: 7	not stimulated vs not stimulated	4	21

The comparisons between the WTs and their corresponding PhoPR KOs under different conditions showed that many more genes were activated under the conditions hypothesized to stimulate PhoR (Figure 41) than under the conditions where PhoR was not expected to be stimulated (Figure 41 A, B, C) (Appendix table 3.6). The only overlap between the activated genes in all four comparisons were *phoP* and *phoR* (Figure 42), whilst seven genes were inactivated under all the conditions when the WTs and their corresponding KOs were compared (Figure 42). These seven genes were phage related (MAB_1725c, MAB_1774, MAB_1775, MAB_1782, MAB_1796, MAB_1801).

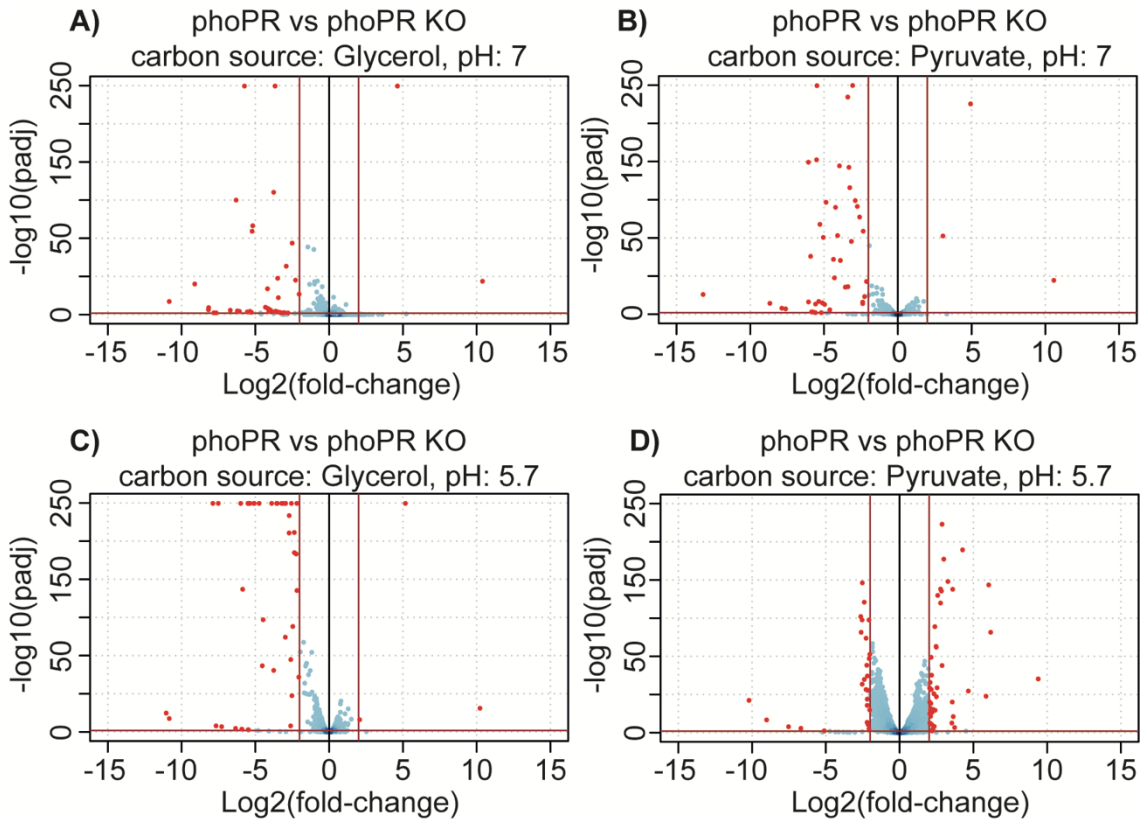


Figure 41: A distinct gene expression pattern was observed when phoPR was grown with pyruvate as a sole carbon source and at a pH of 5.7

Volcano plots showing the significantly differentially expressed (red points) genes between *M. a. massiliense* reference genome CIP108297 and the *M. a. massiliense* CIP108297 phoPR knockout (KO) under the following conditions: A) sole carbon source Glycerol and pH 7, B) sole carbon source Pyruvate and pH 7 C) sole carbon source Glycerol and pH 5.7 and D) sole carbon source pyruvate and pH 5.7. Under the conditions examined in A, B and C, where it was hypothesized that phoR was not stimulated, the majority of the significantly differentially expressed genes were inactivated (< -2 log2fold-change). Contrastingly, under the conditions compared in D, where phoR was hypothesized to be stimulated, more genes were activated than inactivated.

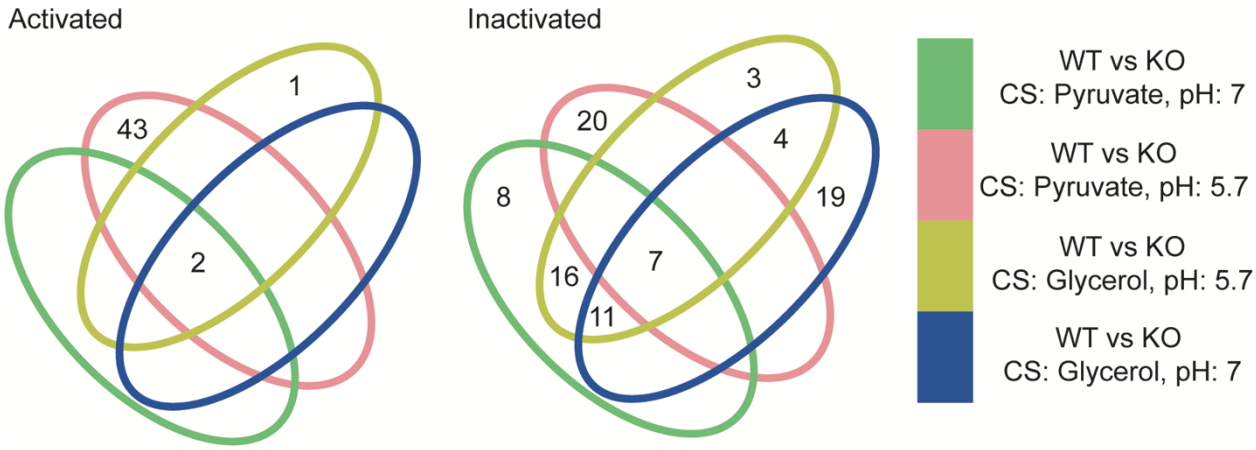


Figure 42: Minimal overlap between differentially expressed genes under conditions where *phoR* was hypothesised to be stimulated and where it was not.

Venn diagrams showing the overlapping DE genes identified from the comparisons between the *M. a. massiliense* CIP108207 WT and their corresponding PhoPR KOs. Apart from *phoP* and *phoR* there were no genes activated under all the conditions tested, whilst there were 7 genes inactivated under all conditions. The limited overlap between the DE patterns produced when the WT and PhoPR KO were compared having been grown at pH 5.7 and with a sole carbon source of Pyruvate (pink) to the other conditions suggested that potentially these conditions had resulted in the stimulation of PhoR.

The distinctive DE pattern observed under the conditions where PhoR was hypothesized to be stimulated suggested that potentially the DE genes could be part of the PhoPR regulon in the MABSC. The functions of the 43 genes activated when the WT was compared to the PhoPR KO under PhoR stimulatory conditions showed some functional similarity to some of the genes known to be regulated by the PhoPR TCS in *M. tuberculosis* H37Rv (Figure 43) (315, 332). Amongst the 43 genes activated were the DosS/R TCS (MAB_3890c, MAB_3891c) and three of the four genes (MAB_2489, MAB_3903, MAB_3904) that have been shown previously to be under the control of the DosS/R TCS in *M. a. abscessus* ATCC19977 (181, 333). There was also activation of genes involved in the metabolism lipids (MAB_3486, MAB_3487) and fatty acids (MAB_3354) and further genes associated with the hypoxic response (MAB_1130, MAB_0869c). Several genes involved in the metabolism of sulfur were also activated (MAB_1652, MAB_1653, MAB_2217), which, whilst they haven't been associated with the PhoPR regulon in other species previously, have been shown to be activated in nutrient limited and oxidative stress conditions (334). Whilst there is some functional similarity between the potential MABSC PhoPR regulon presented here and that of *M. tuberculosis* H37Rv, the only DE genes orthologous to genes in the *M. tuberculosis* H37Rv PhoPR regulon were *phoP* and *phoR*, the sensor histidine kinase component, *dosS*, of the DosS/R TCS and a gene involved in the initial hypoxic response (MAB_3903) (appendix table 3.6) (230, 315).

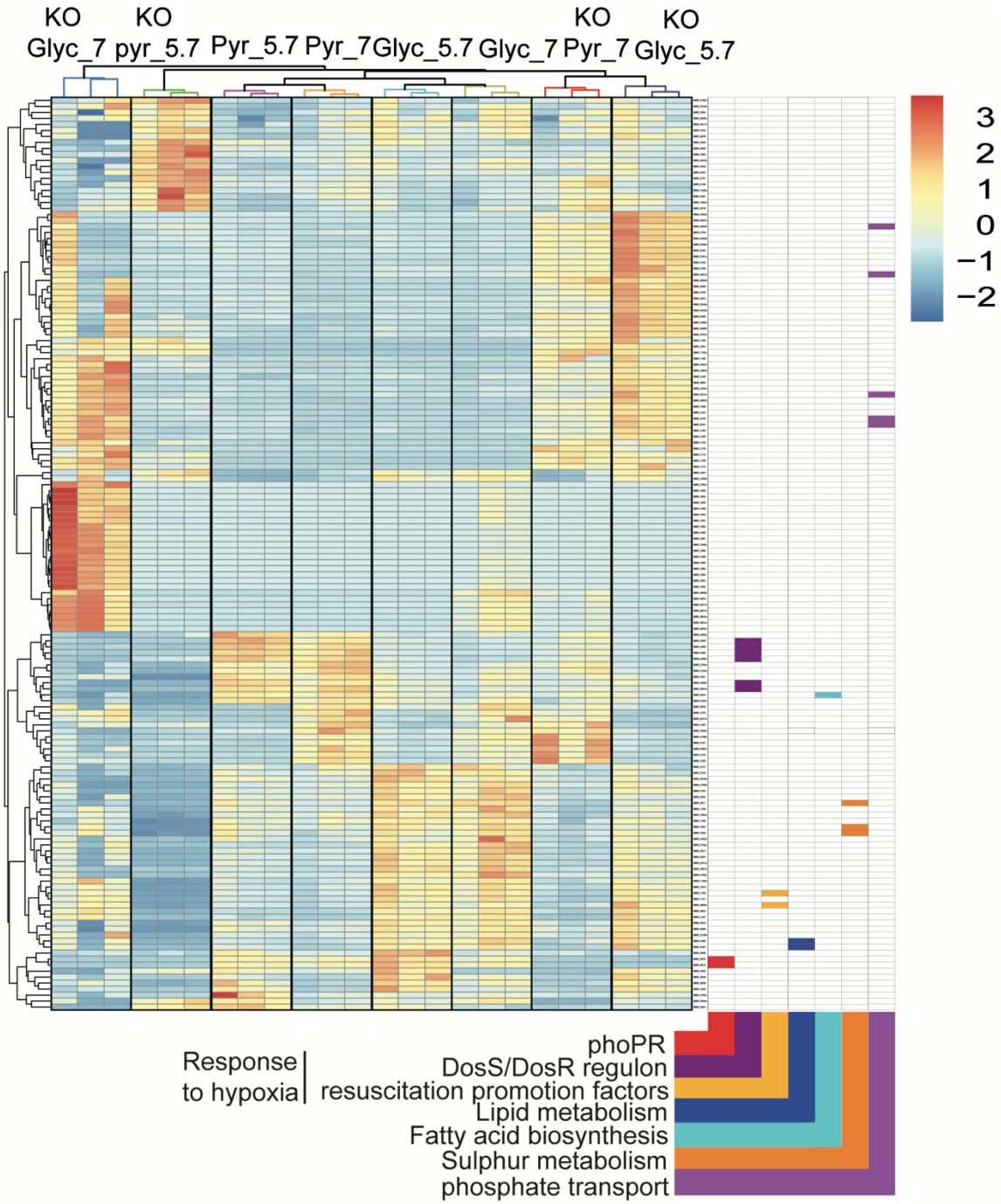


Figure 43: Overlap between the MABSC and *M. tuberculosis* phoPR regulons

Heatmap of the normalised read counts showing the difference in read depth for the 151 differentially expressed genes identified from all the comparisons performed. Highlighted are some of the functions of the DE genes identified from the comparison between Pyr_5.7 and KO_pyr_5.7, which have some similarity to the genes under the control of the *M. tuberculosis* H37Rv phoPR regulon.

The inactivation of 20 genes when PhoR was believed to be stimulated suggested that the PhoP also acted as a repressor. Amongst the 20 genes was a gene upregulated by PhoP in

M. tuberculosis H37Rv, MAB_3100, an L-alanine dehydrogenase, which is orthologous to Rv2780 (315, 332). Inactivation of a gene involved in fatty acid biosynthesis (MAB_0759c), as well as further metabolism associated genes (MAB_1628c, MAB_3840, MAB_4198c) and several membrane associated proteins (MAB_1989c, MAB_4721, MAB_4517c, MAB_1539c) were also observed (appendix table 3.6). Overall the DE expression profile under these conditions provides some evidence to suggest that the PhoPR regulon in the MABSC could potentially be controlling genes involved in both the metabolic adaptation of the organism to enable survival in the phagosome as well as aiding the organisms survival by activating stress response operons.

Contrastingly, the DE patterns produced by comparing the WTs and PhoPRs KO when they were grown under conditions where PhoR was not thought to be stimulated showed that just one gene, other than *phoP* and *phoR*, was activated and only under one condition (Figure 41, 42). However, 61 genes were inactivated under these conditions suggesting that phoPR TCS was having some regulatory effect. Eleven genes were inactivated under all the conditions where PhoR was not thought to be stimulated (Appendix table 3.6). Amongst these 11 genes were two genes involved in phosphate transport (MAB_0746, MAB_0747), two genes involved in the transport of 2-aminoethylphosphonate (MAB_1501 and MAB_1502) and potentially two genes involved in the metabolism of this substrate (MAB_1499 and MAB_1500). The overlap of genes inactivated under all the conditions where PhoR was thought not to be stimulated either suggested i) that PhoR was actually being stimulated and that stimulation of PhoR under these three conditions (Table 13) resulted in the inactivation of these genes, ii) that unphosphorylated PhoP was potentially able to block the transcription of these genes by other regulatory elements or iii) it potentially suggested that a secondary mutation had occurred resulting in the DE of genes not associated with the PhoPR regulon.

Given that there was evidence that PhoR was stimulated by the environmental cues of a low pH and sole carbon source of Pyruvate and that this led to the expression of genes potentially aiding its survival in the phagosome (Figure 43) and that there was no evidence of such a response caused when the organism was grown under the other conditions, which could be interpreted as PhoR not being stimulated, it was hypothesized that a similar DE pattern would be produced when comparing the DE genes between the PhoR stimulated WT and the WTs in which PhoR was not believed to have been stimulated. Table 13 shows the number of DE genes identified when the gene expression levels between the stimulated PhoR WT were compared to the WTs gene expression levels under conditions not thought to

stimulate PhoR. Only four of the DE genes (MAB_2489, MAB_3903, MAB_3904, MAB_2792c) identified in the stimulated WT versus PhoPR KO analysis were also found to be DE between the stimulated WT and not stimulated WTs (Figure 44, 45). Three of these four genes form part of the DosS/R regulon, although contrastingly to the stimulated PhoR WT versus PhoPR KO DE analysis, the genes encoding the DosS/R TCS were not found to be DE.

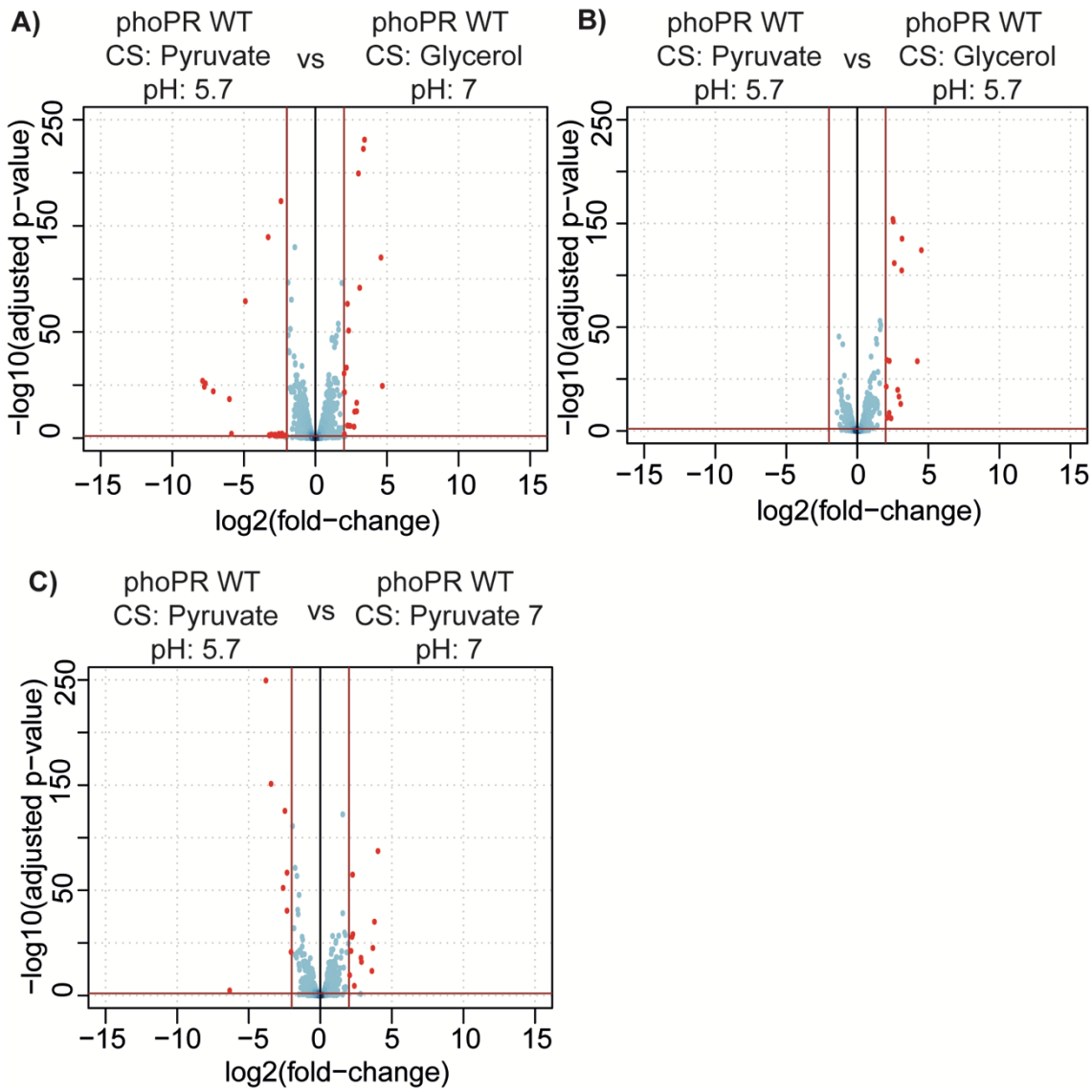


Figure 44: The expression patterns between isolates where phoPR was stimulated and not stimulated did not replicate the expression pattern between the stimulated phoPR WT when it was compared its KO

Volcano plots showing the differentially expressed genes when comparing the expression of *M. a. massiliense* CIP108297 under conditions were it was hypothesized that phoPR would be stimulated vs *M. a. massiliense* CIP108297 grown under conditions where it was hypothesized phoR was not stimulated.

This lack of overlap could have been due to the changes in expression caused by the stimulation of PhoR not being strong enough to be observed without the effect of PhoPR KO or the differing environmental conditions causing other regulatory machinery to interact with PhoPR regulon. However, it was also possible, given the unexpected overlapping DE genes identified between all the WT versus PhoPR KO analyses under the conditions not thought to stimulate PhoR, where no overlap would have been expected given the different environmental conditions and hypothesized lack of stimulation of PhoR, that a secondary mutation was causing this overlap and potentially contaminating the DE analysis.

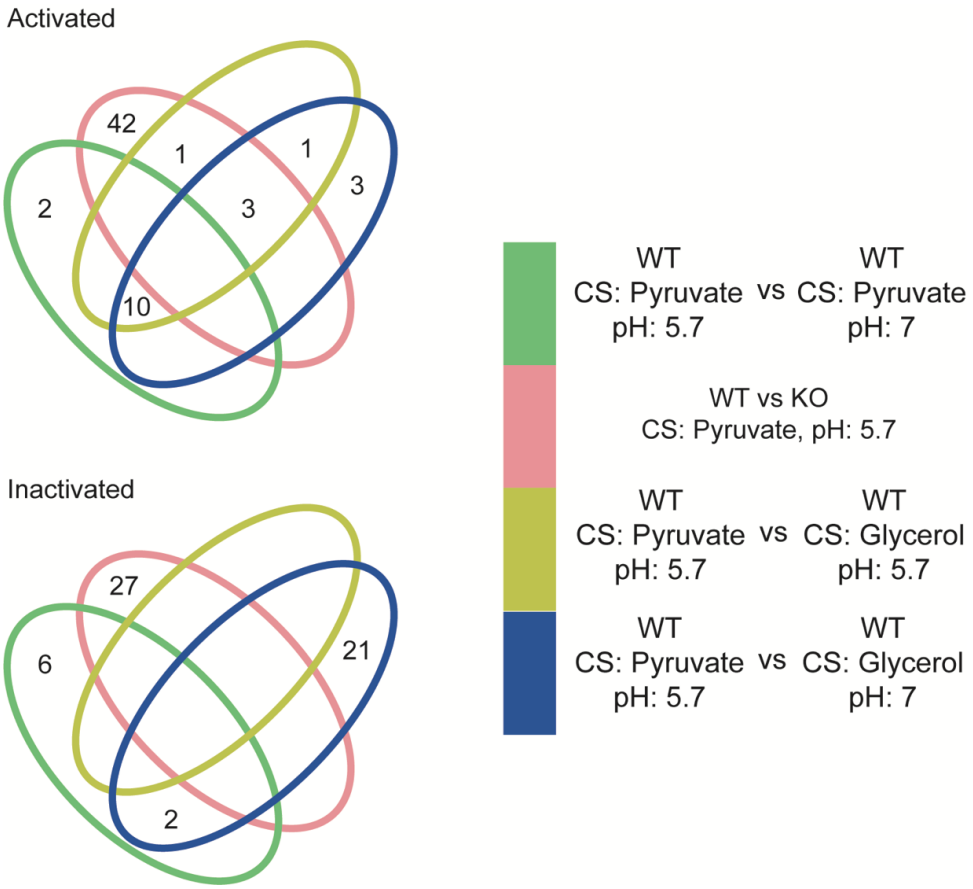


Figure 45: Limited overlap between the DE genes identified between the stimulated and not stimulated PhoPR WTs and the DE genes identified between the stimulated PhoPR vs the PhoPR KO.

Venn diagrams showing the overlap between DE genes identified from the comparisons between the PhoPR stimulated and not stimulated WTs (Green, gold, blue) and the PhoR stimulated WT vs its corresponding PhoPR KO (pink). Overlap between the DE patterns was expected, however, no genes were activated or inactivated across all the comparisons performed. Only four genes identified as possibly part of the PhoPR regulon in the stimulated PhoR versus PhoPR KO DE analysis, three of which were part of the DosS/R regulon, were also found to be DE between the stimulated PhoR and not stimulated WT experiments.

Using WGS of the mutant and parent strains, a secondary mutation, a frameshift in the phosphate transport system permease (*pstA1*) gene, MAB_0748, was identified in the phoPR KO strains. Amongst the inactivated genes identified between WT and phoPR KOs were several genes involved in phosphate transport (MAB_4047c, MAB_4048c, MAB_0746, MAB_0747, MAB_0751c), which could be due to the loss of function of *pstA1* (Figure 43) (Appendix table 3.6). However, the similarity in functions of some of the DE genes identified between the WT and PhoPR KO when PhoR was believed to be stimulated and those known to be controlled by PhoPR in *M. tuberculosis* H37Rv suggested that potentially some of the signal from the activation of PhoR was also being observed.

4.3.6 Further regulators under selection

Three further regulators also accumulated a significant number of nonsynonymous SNPs in parallel in multiple patients. MAB_3036c, a possible ribonucleotide reductase repressor (*nrdR*) and orthologous to *M. tuberculosis* H37Rv gene Rv2718c, accumulated four nonsynonymous SNPs in three patients (Table 12). Whilst Pfam identified the cone domain (amino acids 46-134) characteristic of these regulators, an alignment to its ortholog in *M. tuberculosis* H37Rv, revealed the presence of the other conserved domain, the zinc ribbon domain (Figure 46) (335, 336). Three of the four nonsynonymous changes were accumulated within the cone domain.

MAB_0115c, a conserved hypothetical protein with no domains predicted through Pfam, accumulated five nonsynonymous SNPs in five patients. MAB_0115c has been shown to be orthologous to the *M. tuberculosis* H37Rv gene, *espR* (230). EspR is a key regulator which controls the expression of the virulence associated ESX-1 secretion system responsible for the secretion of the ESAT-6 antigen and furthermore it has also been shown to be expressed in the presence of phoP in *M. tuberculosis* H37Rv (337).

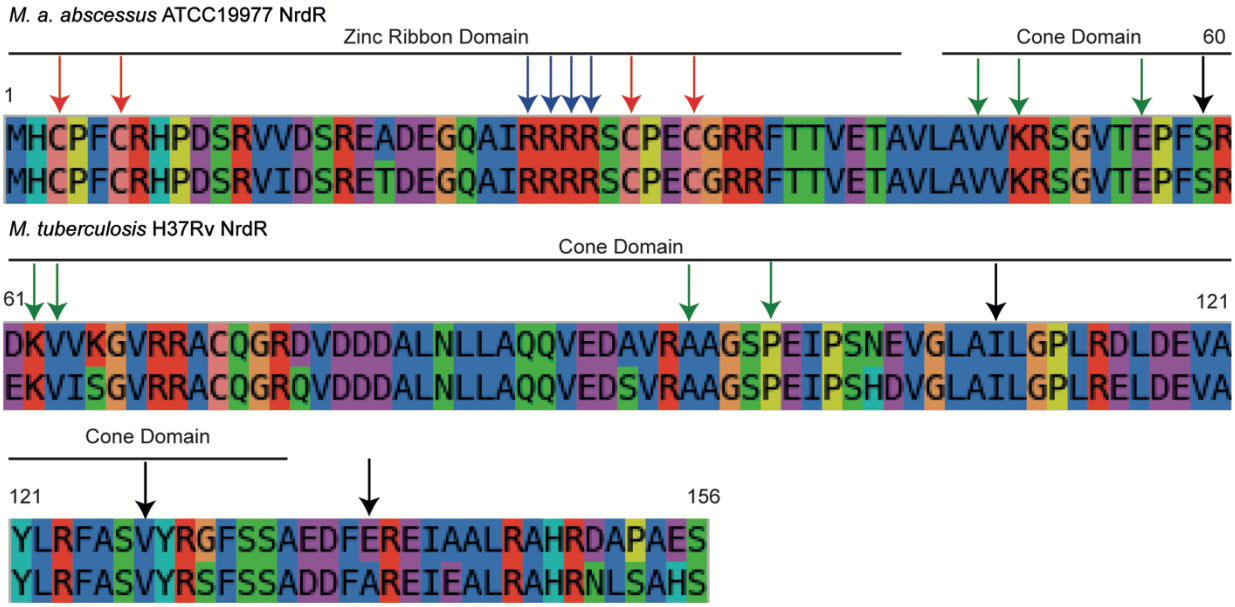


Figure 46: Accumulation of nonsynonymous SNPs in parallel in the cone domain of a ribonucleotide reductase regulator

Protein sequence alignment of MAB_3036c, a ribonucleotide reductase regulator and its *M. tuberculosis* H37Rv ortholog, Rv2718c. The black arrows mark the four nonsynonymous SNPs (E138K,V127A,I109N,S59G) accumulated in parallel in MAB_3036c. The four conserved cysteine residues (red arrows) and four conserved arginine residues (blue arrows) characteristic of a zinc finger domain are indicated. Amino acids residues in the cone domain that have been found to interact with ATP in *E. coli* are indicated by the green arrows. Adapted from figure in Grinberg et al. 2006 (335).

Finally, a *tetR* family regulator, MAB_1881c, accumulated eight nonsynonymous SNPs in seven patients. No ortholog in *M. tuberculosis* H37Rv was predicted. MAB_1881c was predicted to be encoded downstream of a fatty acid biosynthesis operon (MAB_1800c, MAB_1799c, MAB_1798c, MAB_1797c, MAB_1796c, MAB_1795c) and thus could have potentially been regulating this operon. However, MAB_1800c, annotated as a conserved hypothetical protein, was found to encode a *pucR* type regulator helix-turn-helix domain and gene synteny analysis with *M. tuberculosis* H37Rv, showed that the gene order of this operon was conserved between the two species apart from the presence of MAB_1881c and two conserved hypothetical proteins (MAB_1882c, MAB_1883c) downstream of MAB_1881c, which encoded a doxX domain and DM13 domain respectively (Figure 47).

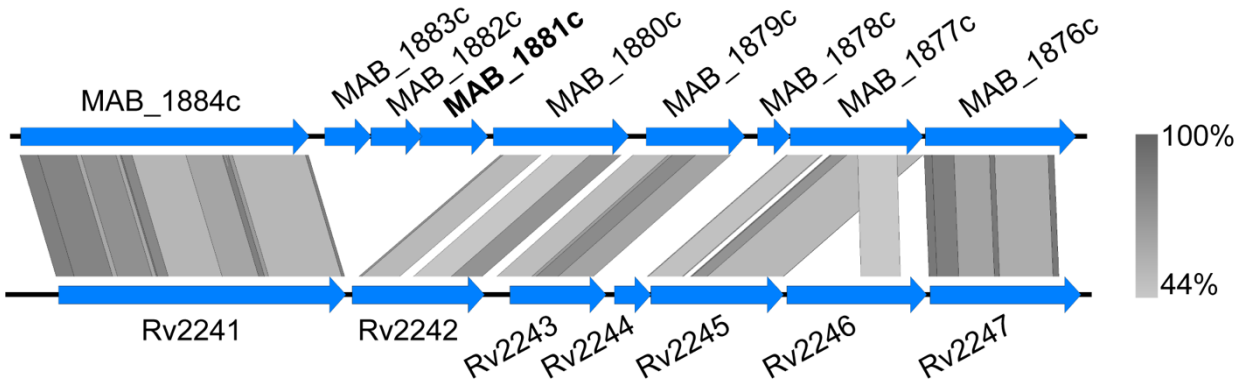


Figure 47: MAB_1881c not regulating FASII operon

Tblastx comparison between the FASII operons in *M. a. abscessus* ATCC19977 (top) and *M. tuberculosis* H37Rv (bottom), regulated by *mabR* (Rv2242 and MAB_1880c). MAB_1881c, the *tetR* family regulator which accumulated seven SNPs in parallel within SNPs forms part of an insertion into *M. a. abscessus* ATCC19977 or deletion in *M. tuberculosis* H37Rv. Figure made with EasyFig (276).

4.3.7 Adaptation of lipids potentially involved in host pathogen interactions

Changes in cell wall composition are also known to have a significant impact on the survival of pathogens within their host. Three genes involved in biosynthesis of cell wall lipids were found to have accumulated a significant number of nonsynonymous and nonsense mutations within patients (Table 12, Figure 35). MAB_0173, which encodes UbiA, a 5-phospho-α-d-ribose-1-diphosphate: decaprenyl-phosphate 5-phosphoribosyltransferase (DPPR synthase), accumulated 11 nonsynonymous SNPs in 11 patients. MAB_0173 encodes a single UbiA domain from amino acids position 32 to 301. All 11 nonsynonymous SNPs were accumulated within this domain, although the MAB_0173 only consists of 305 amino acids. The *M. tuberculosis* H37Rv ortholog of UbiA, Rv3806, catalyzes a step in the synthesis of decaprenyl-phospho-arabinose (DPA) and decapolyprenol-phosphoribose (DPR). DPA is the sugar donor for D-arabinose when in its furanose ring form (Ara*f*) and D-arabinose in turn is a key constituent cell wall lipids arabinogalactan and lipoarabinomannan (LAM) (338).

A second gene involved in the synthesis of LAM, a potential arabinosyltransferase (MAB_0189c), also accumulated a significant number of SNPs over time within patients, with 15 nonsynonymous SNPs observed across 13 patients (Table 12, Figure 35). MAB_0189c has been reported to be orthologous to Rv3793, which encodes the *embC* component of the *embCAB* operon in *M. tuberculosis* H37Rv (230). In *M. tuberculosis* H37Rv, EmbA and EmbB play roles in arabinogalactan biosynthesis, whilst EmbC is involved in the biosynthesis

of LAM (338). Two domains were encoded by MAB_0189c, an arabinose transferase domain from amino acid position 31 to 682 and an EmbC C-terminal domain encoded from amino acid positions 716 to 1082. Of the 15 nonsynonymous SNPs accumulated in parallel by MAB_0189c, 11 fell within the arabinose transferase domain, three fell within the EmbC C-terminal domain and one was not acquired within a domain. Given that 73% (11/15) of the nonsynonymous SNPs fell within the arabinose transferase domain it suggested that the within host selection pressure may be specifically selecting for variants within this domain. However, after randomly introducing 15 variants into the *embC* sequence 1000 times to determine the distribution of nonsynonymous SNPs that would be expected by chance, the observed distribution was not found to be significant (P-value: 0.6999).

MAB_1915, which accumulated 5 nonsynonymous SNPs in parallel, encodes a fatty acid CoA-ligase (FabD). A single AMP binding domain, characteristic of long-chain fatty acid CoA ligases, is encoded by MAB_1915 from amino acid 27 to 490. Three of the five nonsynonymous SNPs accumulated in parallel in this gene were acquired within this domain. However, this gene did not fall within the FasII operon (Figure 46), and nor was it found to be orthologous to a FabD family protein within *M. tuberculosis* H37Rv. The genes upstream of MAB_1915 included a adenylate cyclase (MAB_1914c) and a conserved hypothetical protein (MAB_1913), whilst downstream the genes included a 3-methyl-2-oxobutanoate hydroxymethyltransferase (MAB_1916c) and a conserved hypothetical protein (MAB_1917).

4.3.8 Accumulation of mutations in antibiotic resistance genes

Adaptation to antibiotic resistance is also a key survival mechanism used by pathogenic bacteria and nonsynonymous mutations were accumulated by multiple patients in four known MABSC antibiotic resistance genes. Three nonsynonymous SNPs were accumulated in the *erm*(41) gene in three patients infected with *M. a. abscessus* lineages. The truncation of this gene in *M. a. massiliense* subspecies isolates results in infections caused by this subspecies being susceptible to macrolide treatment, whereas *M. a. abscessus* and *M. a. bolletii* subspecies both have full length *erm*(41) genes which causes inducible resistance to macrolides (77, 339). However, the transition of a T-to-C at position 28 leads to the loss of function of *erm*(41) and subsequent macrolide susceptibility (142). None of the three mutations observed longitudinally occurred at this position (Table 14).

Macrolide resistance can also be acquired by mutations in the 23s rRNA gene in both isolates with a functional *erm*(41) gene and those without (340). Sixteen mutations were observed over time within patients in the 23s rRNA gene. Six were accumulated at position

2270 (equivalent *E.coli* number 2058) with the substitution of A to T seen once, A to C seen twice and A to G seen three times. Four A to G substitutions occurred at position 2271 (equivalent *E.coli* number 2059). Other than A2270T, all the point mutations observed at these two positions (2270 and 2271) have been shown experimentally to cause clarithromycin resistance (313, 340, 341). The point mutation, T2823C, has also been observed previously (341). The remaining five mutations occur at positions that have not, as of yet, been associated with antibiotic resistance (Table 14).

Table 14: Potential antibiotic resistance mutations acquired over time within the host

Gene	Product	No. SNPs	No. unique SNPs	No. patients	Patients	SNPs
MAB_r5051	16S ribosomal RNA	8	4	8	AHL_C_ABSS,AUS_B_ABSS, AUS_I_ABSS,BIR_Z_ABSS, PAP_012_ABSS, SMRL_AT_clone1_MASS, SMRL_BW_ABSS, SMRL_CG_ABSS	**G100T**,**T1086C**,**T1479C**, A1375G,A1375G,A1375G, A1375G,A1375G
MAB_r5052	23S ribosomal RNA	16	10	13	AHL_C_ABSS,AUS_AD_ABSS, AUS_AP_MASS,AUS_J_ABSS, BIR_W_ABSS,BIR_W_ABSS, RHS_M_MASS,RHS_O_ABSS, RHS_O_ABSS,SMRL_BD_ABSS, SMRL_CG_ABSS, SMRL_CG_ABSS, SMRL_CV_ABSS, SMRL_R_MASS, UNC_H_ABSS,UNC_S_ABSS	G768C,A841G,C1829T, **A2270T**,A2270G, A2270C, A2270G,A2270G,A2270C, A2271G,A2271G,A2271G, A2271G,**G2281A**, **T2758G**,T2823C
MAB_2297	Erm(41)	3	3	3	AUS_J_ABSS,RHS_O_ABSS, SMRL_BN_ABSS	P140S,P140R,W10R
MAB_4532c	Eis2	4	4	2	AUS_AD_ABSS,PAP_033_ABSS, PAP_033_ABSS,PAP_033_ABSS	Y227C,C294R,L282P, V77A

MAB_4532c, which encodes an enhanced intracellular survival protein (Eis2), accumulated four nonsynonymous SNPs in two patients. The alignment, constructed using muscle, of the amino acid sequences of MAB_4532c and its *M. tuberculosis* ortholog Rv2416c showed they shared an AAI of 33%. One of the nonsynonymous mutations occurred in the 10th Beta pleated sheet, Y227C, whilst, a further nonsynonymous mutation occurs in the 13th Beta pleated sheet, C294R, in very close proximity to residues which both form and line the aminoglycoside binding pocket (342). L282P occurs at an amino acid residue conserved across the three *eis* homologues compared in (342). The amino acid change, V77A,

occurred in the 4th Beta pleated sheet, flanked by sites which are conserved and in the vicinity of amino acid residues responsible for forming the aminoglycoside binding site (342).

Resistance to aminoglycosides can also be acquired by the MABSC through point mutations in the 16s rRNA gene (156, 343). Eight point mutations were accumulated longitudinally in this gene in eight patients, with five of the eight mutations being a substitution of A to G at position 1375 (equivalent to *E.coli* no. 1408), a point mutation that has previously been associated with aminoglycoside resistance in the MABSC (Table 14) (156, 343). The remaining three point mutations: T1479C, T1086C and G100T have not been reported previously in the literature, to my knowledge.

4.4 Discussion

Identifying genes evolving in parallel in multiple patients is a method that has been applied to multiple CF pathogens to successfully identify novel virulence factors and potential drug targets (40, 182, 307-309, 311, 312). This approach had only been applied on a small scale to examine the within patient evolution of the MABSC (183). The MABSC global population dataset included multiple isolates for 182 patients once isolates potentially acquired by transmission or representative co-infection had been removed (Appendix Figure 3.1). The aim of this chapter was to look for convergent evolution occurring between these MABSC lineages evolving independently in individual patients in order to uncover novel virulence factors which could increase our understanding of the pathogenesis of the MABSC and potentially identify novel drug targets.

The acquisition of mutations over time within patients was found to be common with lineages from 75% of the patients investigated acquiring at least one SNP. The majority of patients appeared to be acquiring mutations in a clock like manner, although the level of variation from one patient stood out. The isolates obtained from SMRL_AT accumulated a level of variation that suggested that the lineage had acquired a hypermutator phenotype (Figure 34A). An indel, leading to a frameshift in the endonuclease III (nth) gene, an enzyme that catalyzes two steps in the base excision repair pathway, was subsequently identified in the isolates from SMRL_AT (323, 324). Mutations in this gene in *E. coli* have been shown to infer a weak hypermutator phenotype (323, 324). The acquisition of a hypermutator could potentially have enabled this MABSC lineage to adapt to the CF lung quicker, as the increased substitution rate increases the chance of acquiring a mutation in a beneficial gene, such as genes involved in pathogenicity or antibiotic resistance, as has been observed in hypermutating lineages of other CF pathogens (344, 345). There is evidence to suggest that

this has occurred in the lineage infecting SMRL_AT as nonsynonymous SNPs were accumulated over time in the potentially pathoadaptive genes *phoR* and *fabD*, as well the antibiotic resistance associated 16s rRNA gene (Table 12,Figure 35). On the other hand, some lineages from other patients without a hypermutator phenotype acquired more nonsynonymous SNPs than the lineage from SMRL_AT in potentially pathoadaptive genes, which could be due to a difference in sampling time, but could also reflect the fact that whilst a hypermutator phenotype initially provides an advantage, it also increases the chances of the bacteria acquiring a lethal mutation (346, 347).

Naturally occurring hypermutators have not been described before in the literature for Mycobacteria, however, a hypermutator in *M. abscessus* has been described previously (Josephine Bryant, unpublished work). The hypermutator phenotype in this patient was due to a premature stop codon in the uracil DNA glycosylase (*udg*) gene (MAB_3283c), which removes uracil that has been mis incorporated into DNA due to deamination of cytosine (348). Although the isolates from this patient, PAP_033, were included in this analysis and were found to have accumulated a premature stop codon in MAB_3283c, the number of variants accumulated, 15 in 1101 days (3 years and 5 days), did not stand out as greater than expected. The lack of evidence in this analysis for PAP_033 being a hypermutator is likely to be due to only consensus SNPs and not minority variants being analysed (Josephine Bryant, unpublished work). This suggests that it is possible that there are more patients whose isolates have acquired a hypermutator phenotype which have not been detected through this analysis, and that hypermutation could be contributing to greater extent than is currently evident to the adaptation of the MABSC to the CF lung.

Further support for the presence of further hypermutators in this MABSC dataset was provided by the substitution rate estimated through this analysis. An average 3.1 SNPs were estimated to become fixed per genome per year which is within the 95% confidence interval for a previously estimated *M. a. abscessus* substitution rate (1.8 substitutions per genome per year; CI: 0.8-3.3) (70). However, this is still higher than expected. One explanation could be that further hypermutators are present in the dataset but haven't been detected. The relatively low R^2 value also suggests that potentially other variables are effecting the correlation. These could be biological variables such as the selection pressure being applied by the individual hosts differing through a combination of factors such as the host's immune system, co-infection with other CF pathogens or antibiotic treatment being taken over the time in which the samples were collected, or they could be methodological. Methodological reasons that could be influencing the correlation include the time span over which the

isolates were collected not reflecting the length of infection, true variants not being called due to not meeting the variant calling criteria resulting in false negatives, isolates with a lower than expected number of SNPs could have acquired variation through other means, such as indels, which were not investigated in this analysis, or it could be that variants in the accessory genome of these isolates were missed due to all the isolates being mapped to *M. a. abscessus* ATCC19977.

Parallel evolution of MABSC lineages was detected between patients, with 17 genes accumulating a greater number of nonsynonymous SNPs than would have been expected by chance in multiple patients (Table 12, Figure 35). Three of the genes, *embC*, *phoR* and *fabD* were also found to have accumulated nonsynonymous SNPs over time within a single patient in Kreutzfeldt and colleagues analysis (183). There is also considerable overlap between the candidate genes identified here and those identified in analysis done previously on a subset (n=31) of the patients used in this analysis, although using a different statistical approach (Josephine Bryant, unpublished work).

Nonsynonymous SNPs were accumulated in these 17 genes by 57 patients, just 31% of the patients investigated. Whilst this could be due to some patients not having been sampled over a long enough period of time for SNPs to have become fixed in the population, it could also suggest that there are other ways in which the MABSC is adapting to the lung and that the signal for this method of adaptation was not strong enough to be detected within this dataset. However, the signal that was detected in this dataset, through the identification of genes evolving in parallel, has highlighted a group of genes with potential roles in the adaptation of the MABSC to the lung.

Functional enrichment analysis carried out using DAVID and the String database and by searching the KEGG database, failed to highlight any significantly enriched functions or pathways amongst the candidate genes. This could be due to being underpowered and in the case of the pathway enrichment analysis could be due to the lack of knowledge about which genes are involved in which pathways encoded by the MABSC. Investigating the candidate genes functions manually showed that regulation was the most common function (7/17) and that genes involved in similar biological processes were present (Figure 35). Most striking was the presence of four regulators which respond to environmental cues emitted bbeduring the maturation phagosome which when coupled with the accumulation of a significant number of nonsynonymous SNPs in three genes involved in cell wall biosynthesis, which is also at the interface of host-pathogen interactions in the phagosome, suggests that

the within host selection pressure is selecting for changes aiding survival within the phagosome (329, 331, 338, 349-352).

The four regulators responding to environmental cues that accumulated a significant number of nonsynonymous SNPs in parallel were IdeR, PhoR, CRP and WhiB1 (Table 12). For three of the regulators, IdeR, CRP and WhiB1 the stimulant they are sensing is known, with IdeR responding to levels of Fe^{2+}, CRP responding to levels of cAMP, which are increased when mycobacteria are under stress conditions, and WhiB1 responding to the presence of nitric oxide (NO) (325, 329, 351, 353). However, the stimulant that causes PhoR to phosphorylate PhoP is less well understood, with a low pH currently believed to be the environmental cue, although dependent on which carbon source is available from the host (331). All these signals are features of the phagosomal environment and have been shown in *M. tuberculosis* H37Rv to act as triggers for the reprogramming of *M. tuberculosis* H37Rv regulatory networks to enable survival in the macrophage (354). Within the CF lung, these are also features of the extracellular environment, with the pH of CF sputum estimated to be 5.2 and areas with hypoxic conditions also created due to the formation of mucus plugs and the presence of aerobic bacteria (26, 32, 355). Consequently, these regulators could also be contributing to the extracellular survival of the MABSC.

In *M. tuberculosis* H37Rv, these four genes have been shown to have overlapping regulons and to potentially be involved in the regulation of one another, resulting in a recent review presenting these four genes as part of a central regulatory cascade that sets in motion the key virulence strategies utilized by *M. tuberculosis* to survive and persist within the host (356, 357). Therefore, it is tempting to speculate, given the selection for variants in four of the regulators/sensors in this cascade, all of which are orthologous to their *M. tuberculosis* H37Rv counterparts, that a similar regulatory control hub is playing a role in the survival of the MABSC within the host. Further research is required to establish the genes controlled by these regulators, particularly in the case of the PhoPR TCS and CRP and WhiB1 regulators, although an initial attempt to determine the PhoPR regulon in the MABSC was carried out in this analysis. Further research is also required to determine the extent of the interaction between the regulators, both in terms of overlapping regulons and in terms of whether they are potentially influencing the regulation of one another.

The position of the variants accumulated could potentially suggest how these regulators are adapting in order to provide a selective advantage to the MABSC. However, the definitive functional impact of the nonsynonymous SNPs acquired by the four regulators cannot be

proved without experimental analysis. Preliminary sequence analysis showed that all four of the regulators accumulated at least one variant in either domains or at positions with known functional roles in their corresponding *M. tuberculosis* H37Rv orthologs, which could be seen as evidence that the mutations accumulated within the host are changing the function of the protein (Figure 36, 37, 38, 40).

IdeR accumulated a variant in the predicted DNA binding motif, although not at a position known to make contact with the DNA (green arrow, Figure 36), which could potentially suggest that the IdeR DNA binding affinity might be affected as opposed to the complete loss of its DNA binding capability (325). Three of the other variants detected in this gene also fall near the DNA binding motif and therefore could also potentially be affecting the DNA binding capabilities of IdeR (Figure 36). IdeR, of which there is only a single copy in *M. a. abscessus* ATCC19977 and *M. tuberculosis* H37Rv, has been shown to be essential in *M. tuberculosis* as it regulates the genes responsible for iron homeostasis; both too little and too much Fe^{2+} is detrimental to the organism's survival (350). Further research is required to determine the functional impact of the nonsynonymous SNPs accumulated by IdeR.

One of the seven variants acquired by WhiB1 occurred at a position W49R, which in the structure of its *M. tuberculosis* H37Rv ortholog forms the mouth of a channel that allows NO access to the [4Fe-4S] cluster, destabilization of which results in WhiB1 being able to bind DNA (Figure 38) (329). However, whether this amino acid change makes the [4Fe-4S] cluster more accessible, which could be interpreted as beneficial as it might make WhiB1 more sensitive to NO resulting in a faster transcription response by the MABSC to oxidative stress conditions, or whether alternatively it does the opposite and blocks the channel resulting in a slower or complete lack of a response, is unclear. However, given that this change is from a non-polar hydrophobic amino acid (tryptophan) to a polar basic amino acid (arginine), a structural change seems probable. Only one variant occurred at a position with functional information available and thus it is not possible to determine, without further research, the impact of the nonsynonymous SNPs acquired by WhiB1 (329).

Three of the seven nonsynonymous SNPs acquired by CRP occurred at sites directly interacting with cAMP, R89Q twice and E80G (Figure 37) (328, 358). This suggests that the ability of CRP to bind cAMP might be affected, particularly as both amino acid changes result in an amino acid with a charged side chain being replaced by an amino acid with an uncharged side chain, although in both cases the polarity of the amino acid remains the same. However, the exact impact of the nonsynonymous SNPs on CRP cannot be

determined without experimental analysis. But the accumulation of variants in this transcription factor suggest it is regulating genes associated with virulence. It's ortholog in *M. tuberculosis*, Rv3676, has been shown to be required for virulence and amongst its putative regulon was WhiB1 which was also found to accumulate a significant number of nonsynonymous SNPs within the host (357).

PhoR provided the strongest evidence that the within host selection pressure was selecting for variants in specific domain, with 70% of the nonsynonymous SNPs being accumulated in the sensor loop, a pattern not expected by chance (Figure 39, 40). The PhoPR TCS is a key virulence factor in *M. tuberculosis*. It has been shown to regulate complex lipid biosynthesis, genes involved in the response to hypoxia, the secretion of the immunogenic protein ESAT-6 and activation of the DosS/R regulon (315, 359, 360). It is tempting to hypothesize that the accumulation of nonsynonymous SNPs in the sensor loop is due to PhoR adapting to the environmental cue it is detecting within the host, potentially either due to being exposed to differing concentrations of the stimulant or even potentially a different stimulant to what it senses in its natural environment. This would fit in with the idea of the within host selection pressure potentially selecting for variants in these regulators that adjust the interaction between either the regulator and its environmental cues or the regulator and its DNA binding site, resulting in a faster response to the changing environment in the host.

However, intriguingly, nonsynonymous variants accumulated in the PhoR sensor loop have been observed previously in *M. tuberculosis* species complex organisms, with a nonsynonymous SNP in the sensor loop of PhoR occurring in the last common ancestor of *M. africanum* L5 and L6 and all animal adapted *M. tuberculosis* complex lineages (361). This variant downregulates the PhoPR regulon and reduces virulence, although compensatory mechanisms which partially restored the expression of the PhoPR regulon have been identified in some of these lineages (361). The impact of the sensor loop mutation, which when it was introduced into *M. tuberculosis* H37Rv resulted in a reduced PhoPR regulon and reduced virulence, potentially suggests that the accumulation of variants in the sensor loop in the MABSC could be causing a reduction in virulence, although this is based on the assumption that the *M. tuberculosis* and MABSC PhoPR regulons are similar (361). In order to discover whether the *M. tuberculosis* and MABSC PhoPR regulon were similar, RNA-seq analysis was performed.

The initial RNA-seq results comparing the PhoPR WT grown under the conditions hypothesized by Baker et al. (2014) to stimulate the PhoPR dependent response to a low pH

suggested that there was some overlap between the two regulons, with the DosS/R regulon, resuscitation promotion factors (rpfs) and genes involved in lipid metabolism activated in the WT in comparison to the KO (Figure 43) (315, 331, 360). The striking difference between DE pattern produced under these conditions and the DE pattern produced when the WT and KOs under the alternative conditions which were not thought to activate the PhoPR regulon, suggested that this was the PhoPR regulon (Figure 41, 42). However, comparing the WTs under PhoR stimulated and not stimulated conditions, failed to reproduce these results (Figure 44, 45). Plausible explanations for this could be that the signal was not strong enough to detect, or other regulators which respond to the differing environmental conditions were also influencing the control of PhoPR regulated genes. However, the identification of a secondary mutation knocking out the *pstA* gene, which forms part of a phosphate transmembrane channel which transports inorganic phosphate into the cell as well as repressing the SenX3 sensor component of the SenX3/RegX3 TCS in the presence of high phosphate concentrations, explained why SenX3/RegX3 TCS and several other phosphate transport associated genes appeared to be being suppressed by PhoPR in the WT vs KO comparisons (Figure 43) (362). Although, these genes were only evident under the conditions where PhoPR was not stimulated, which could be due to PhoPR also inducing the expression of these genes, one of which, *phoY*, is induced by PhoPR in *M. tuberculosis* H37Rv and therefore these genes wouldn't be seen as DE between the stimulated WT and KO given the KO also included the loss of function of *pstA* (332). This secondary mutation complicates the interpretation of the results as the DE genes identified as the possible PhoPR regulon, also include those DE expressed due to the indel in *pstA*. Consequently, further research is required to decipher the genes controlled by the MABCS PhoPR regulon.

However, if the assumption that the PhoPR regulons are similar is correct, as the RNA-seq analysis results provide some imperfect support for, this suggests that PhoPR regulon is controlling virulence related genes as well as providing evidence that the PhoPR regulon is being activated by conditions similar to the CF lung, both in the sputum and the macrophage. Consequently, depending on the impact of the nonsynonymous SNPs on the function of PhoR, the variants accumulated in parallel could either cause a similar effect to those in the sensor loop in *M. tuberculosis* complex species, reducing the expression of the PhoPR regulon and reducing virulence (361). Although, it should be noted that a secondary mutation could potentially be recovering some of the PhoR function, as has been seen in an *M. bovis* outbreak strain (361). Whilst this does not support the hypothesis proposed earlier that the selection for variants in this domain is increasing its sensitivity to its environmental cue, resulting in key virulence regulons being activated faster, it could potentially suggest that

these variants are being selected for because they result in reduced virulence, which been shown to help the persistence of bacteria causing chronic lung infections, including other CF pathogens (363, 364).

A significant amount of experimental analysis is required to determine the impact of these nonsynonymous SNPs on their respective protein functions, as well as RNA-seq and/or CHIP-seq analysis, if the hypothesis that a central regulatory control hub is coordinating MABSC response to within host environmental cues in a similar way to *M. tuberculosis* is to be tested. Whilst the generation of a PhoPR KO without a secondary mutation to determine which genes are regulated by PhoPR is also necessary. The final step to decipher the impact of the PhoR sensor loop SNPs would be to introduce the observed variants individually into a reference *phoR* gene and perform RNA-seq to determine their impact on the regulon.

Three further regulators accumulated a significant number of nonsynonymous SNPs in parallel (Figure 35), and furthermore, two of the *M. tuberculosis* H37Rv orthologs of these regulators, Rv3849 and Rv2718, are also involved in the regulatory reprogramming of *M. tuberculosis* H37Rv to enable survival in the macrophage (356). One of the regulators, MAB_0115c, was orthologous to *espR* (Rv3849) which in *M. tuberculosis* H37Rv, is not only regulated by PhoP, but also regulates, amongst other virulence associated genes, the genes encoding the ESX-1 secretion system which is responsible for transporting a key virulence factors such as ESAT-6 protein which is involved in phagosomal rupture and immune modulation (337, 365). Given that the MABSC only encodes ESX-3 and ESX-4, EspR in the MABSC cannot be regulating the expression of ESX-1 or ESAT-6 which potentially explains why, in the RNA-seq analysis performed in this analysis, EspR is not found to be activated by PhoP, although this could be due to the secondary mutation (72). However, analysis in *M. tuberculosis* H37Rv has shown that EspR effects the regulation, both positively and negatively, of multiple cell wall associated genes and other potential virulence factors, including those also part of the regulons of PhoP and CRP; this suggests that MAB_0115c could also be regulating genes playing a role in pathogenesis and is further support for within host selection for variants in global regulators with the potential to enact large scale regulatory changes (365).

MAB_3036c encodes a ribonucleotide reductase repressor (*nrdR*). Ribonucleotide reductases (RNRs) perform the critical function of catalysing the breakdown of 5'- di or tri-phosphates into deoxynucleotide triphosphates (dNTPs) which are needed for DNA replication and repair (336). Three of the observed nonsynonymous SNPs fall within the cone

domain (Figure 46), which contains the binding site for either ATP or dATP, which when bound by NrdR maintains the repressor in a structure where it is able to bind DNA, although a more complex allosteric mechanism has also been proposed (275). Thus, it is possible that the three variants that were accumulated in this domain could potentially be affecting the binding of dATP or ATP, and consequently its function. Expression of RNRs is believed to vary in response to changing environmental conditions, such as nutrient and oxidative stresses, with derepression of RNRs aiding the recovery from DNA damage which provides further support for the within host selection for variants in a regulator responding to environmental cues (366). Investigations into the role of NrdR have shown that increased expression of NrdR results in decreased growth and fitness in *E. coli,* whilst over expression of RNRs in *B. subtilis*, when *nrdR* was experimentally KO, has been shown to induce stationary phase adaptive mutagenesis, resulting in the acquisition of mutations in genes that aid the organisms survival (366, 367). Contrastingly depression of RNRs in *M. smegmatis* did not confer a hypermutator phenotype or affect the growth rate (336). This fits with the hypermutator observations from this analysis, where none of the patients (AHL_C, SMRL_CG, SMRL_L) which accumulated a SNP in NrdR appeared to have acquired a hypermutator phenotype, although the number of SNPs accumulated by these patients is quite high, which could suggest that this interpretation is wrong and that a similar scenario as to what is seen in *B. subtilis* is occurring (Appendix table 3.2) (366). Whilst further research is required to determine the impact of the variants on the function of NrdR, if they were to cause the same effect as the reduced growth and fitness phenotype displayed by *E.coli* then this could be evidence of *M. a. abscessus* adapting to persistence, whilst if the variants result in the stationary phase mutagenesis phenotype displayed by *B. subtilis*, it could be evidence of a mechanism used by *M. a. abscessus* to escape host defenses (366, 367).

The final regulator, a *tetR* family regulator, MAB_1881c, is encoded just downstream of the FASII fatty acid biosynthesis pathway, which plays a role in pathogenesis as through this pathway key constituents of the Mycobacterial cell wall are synthesized (368). However, the regulator for this operon, *mabR*, is encoded by the adjacent CDS to the candidate gene, suggesting that it is unlikely that the candidate regulator is influencing this regulon, particularly as this operon is conserved within *M. tuberculosis*, and it is clear that the regulator of interest has either been inserted into *M. a. abscessus* ATCC19977 or deleted from *M. tuberculosis* H37Rv along with two other CDS annotated as conserved hypothetical proteins (Figure 47) (368). Consequently, the function of this regulator remains unclear.

Whilst the MABSC appears to predominantly be using changes in regulation to adapt to the host, there was also evidence of parallel evolution within genes involved in cell wall lipid biosynthesis and evidence of the emergence of antibiotic resistance over time within the host. Cell wall lipids are at the interface of host pathogen interaction and two genes, MAB_0173 (*ubiA*) and MAB_0189c (*embC*) involved in cell wall lipid biosynthesis were found to have accumulated significantly more SNPs than would have been expected by chance (Table 12). The functions of the *M. tuberculosis* H37Rv orthologs of these two genes, Rv3806 and Rv3793 respectively, suggest that there could potentially be overlap in the effects of acquiring mutations in these two genes. UbiA catalyzes a step in the synthesis of the sugar donor, DPA, which donates a sugar to *Araf* (the furanose ring form of D-arabinose), which is a key constituent of both arabinogalactan and LAM, whilst EmbC, after an initial Ara*f* molecule is bound to the lipomannan backbone, continues to extend the chain of Ara*f,* resulting in the formation of LAM (338). A further step results in the addition of mannose residues to LAM, leading to the formation of mannose-lipoarabinomannan, which has been shown in *M. tuberculosis* H37Rv to potentially play a role in halting phagosome maturation (352). This suggests that the mutations acquired longitudinally in these genes could be associated with the interaction between *M. a. abscessus* and the host, although further research is required to determine their function in the MABSC and the consequences for their interaction with the host.

Initially, the accumulation of variants in MAB_1915, a fatty acid CoA-ligase (FabD), suggested it could also be playing role in cell wall biosynthesis. However, it does not fall within the FASII operon in *M. a. abscessus* ATCC19977 (Figure 47), and nor is it orthologous to the *fabD* gene in this operon in *M. tuberculosis* H37Rv, which suggests it is not performing a role in this pathway and consequently it isn't possible to associate the accumulation of variants in this gene with the biosynthesis of the mycolic acids in the mycobacterial cell wall. In addition, the genes flanking MAB_1915, fail to shed light on the possible function of this gene.

Two cell wall biosynthesis genes which might have been expected to have accumulated a significant number of SNPs longitudinally were the genes that encode the GPLs as these are associated with the switch from a smooth to rough morphotype which has been association with virulence (160). Seven and eight nonsynonymous SNPs were accumulated in parallel in the genes, MAB_4098 and MAB_4099 respectively, however, the number of SNPs accumulated by these genes was not found to be significant (Appendix table 3.4, 3.5). This is likely to be due to the length of these genes, and the lack of synonymous SNPs accumulated

in parallel within them, resulting in very a high expected number of nonsynonymous SNPs. Whilst this is a limitation of this method and can result in false negatives, the lengths of these genes are exceptional and thus it is unlikely to be having a significant impact on the results.

Acquisition of antibiotic resistance mutations has been shown to commonly occur within the CF lung given the long courses of antibiotics people with CF undertake. The MABSC is already one of the hardest infections to treat and resistance and adverse reactions to the only treatment protocol of an aminoglycoside, macrolide and one or more parenteral antibiotics are common (55). Thus it was not surprising to find mutations occurring in parallel in antibiotic resistance associated genes[10], with many of the variants detected in 16s rRNA and 23s rRNA genes, causing aminoglycoside and macrolide resistance respectively, having been described previously (Table 14) (313, 340, 341). Although potentially some novel variants causing resistance were identified (Table 14, in bold).

Variants were also detected in MAB_4532c, which encodes Eis2. MAB_4532c, has recently been linked with antibiotic resistance due its upregulation by WhiB7, which is induced by the presence of the antibiotic amikacin (152). Thus the variants accumulated within this gene could be associated with antibiotic resistance. In *M. tuberculosis* H37Rv the *eis* gene, Rv2416c, which was found to be orthologous to MAB_4532c, has been linked with intracellular survival as well as being associated with kanamycin resistance (369, 370). Therefore it is possible that MAB_4532c could play a dual role as well. Further research is required to determine the role MAB_4532c is playing in the adaption of the MABSC to the CF lung.

Interestingly, variants were also detected in *erm(41)*, which causes inducible macrolide resistance, in three patients infected with *M. a. abscessus* lineages which have the full length *erm(41)* gene and consequently are resistant to macrolides, which makes the acquisition of variants in parallel within this gene unexpected (77, 339). However, the fact that two of the three patients, AUS_J and RHS_O, also accumulated a SNP in 23s rRNA gene suggests that potentially these nonsynonymous mutations could be causing the loss of function of *erm(41)*, similarly to the T28C mutation, which would explain the need for a mutation in another macrolide resistance associated gene (77, 339). Furthermore, both these isolates accumulated a SNP in the same position (P140S, P140R), suggesting that this could

[10] The 16s rRNA and 23s rRNA genes were not included in the statistical test for accumulation of a greater than expected number of snps but variants accumulated over time within these genes were recorded.

potentially be a previously undescribed variant that can cause the loss of function of *erm(41)*. It is less clear what effect the variant accumulated in the third patient has, but the lack of a compensatory mutation in a gene conferring macrolide resistance suggests that it is unlikely to cause the loss of function of *erm(41)*.

4.5 Conclusions and Future Directions

Overall this analysis has shown that there is strong evidence of parallel evolution of the MABSC within the host. Through this analysis it was possible to identify genes potentially playing a role in the adaptation of the MABSC to the CF lung. The MABSC was found to be adapting to the host through mechanisms commonly used by other pathogenic microorganism, including other CF pathogens. Evidence of the acquisition of a hypermutator phenotype was discovered, along with the detection of convergent evolution in genes associated with regulation, cell wall lipid biosynthesis and antibiotic resistance related genes.

Particular attention was paid to the regulators, the majority of which responded to environmental cues emitted within the phagosome. Many of these regulators had been associated with the regulation of virulence genes in other pathogenic organisms, suggesting that this analysis has potentially uncovered key regulators of virulence genes in the MABSC. Further research is required to determine what genes are under the control of these regulators through which our understanding the genes involved in the pathogenesis of the MABSC could be greatly increased. The first attempt to characterize the MABSC PhoPR regulon was attempted in this analysis with the results suggesting that it regulates genes with similar functions to those under the control of PhoPR in *M. tuberculosis* H37Rv, although this analysis was confounded by the presence of a secondary mutation.

Through this analysis genes potentially associated with the pathogenesis of the MABSC were uncovered, further research is required to validate the functions of these genes and determine whether any are promising drug targets. However, it is interesting to note that some of the candidates highlighted through this analysis, such as UbiA, and PhoP of PhoPR have been suggested before as good candidates for potential novel drugs or vaccines (371, 372).

5. Investigating the epidemic of *M. a. massiliense* post-surgical wound infections in Brazil

The results reported in this chapter are published in:

Everall, I., Nogueira, C., Bryant, J., Sánchez-Busó, L., Chimara, E., Duarte, R., Ramos, J., Lima, K., Lopes, M., Palaci, M., Kipnis, A., Monego, F., Floto, R., Parkhill, J., Leão, S. and Harris, S. (2017). Genomic epidemiology of a national outbreak of post-surgical *Mycobacterium abscessus* wound infections in Brazil. *Microbial Genomics*, 3(5).

And in the manuscript in preparation:

Vinicius calado Nogueira de Moura, Deepshika Verma, Isobel Everall, Crystal Shanley, Megan Stapleton, Julian Parkhill; R. Andres Floto, Diane J. Ordway, and Mary Jackson (2017). Role of outer membrane porins in the fitness, virulence and biocide resistance of *Mycobacterium abscessus*. (Manuscript in preparation)

Statement of contribution:

I carried out all bioinformatics analyses reported in this chapter. The DNA for the isolates sequenced from Brazil was extracted by Vinicius calado Nogueira de Moura. DNA sequencing was carried out by the high throughput sequencing pipeline at the Wellcome Sanger Institute. The glutaraldehyde tolerance assays were carried out by Vinicius calado Nogueira de Moura. These projects were supervised by Julian Parkhill, Andres Floto, Simon Harris, Sylvia Leão and Mary Jackson. All authors contributed to the interpretation of the results.

5.1 Introduction

All three subspecies of MABSC are capable of causing opportunistic infections. Most commonly they cause pulmonary infections in those with underlying lung conditions and skin and soft tissue infections (SSTIs). Outbreaks of both pulmonary infections and SSTIs caused by the MABSC have been reported for decades. The majority of these outbreaks have been point source and on a small scale. However, there are a few exceptions, the outbreaks of pulmonary infections in CF patients in Papworth hospital in the UK and Seattle in the U.S, which were caused by indirect person to person transmission and the epidemic of post-surgical wound infections in Brazil (70, 119, 130, 373). Whilst the outbreaks of pulmonary infections in the CF centers in the U.S and the U.K have been thoroughly investigated through WGS, leading to significant breakthroughs in the understanding of the MABSC, only a few isolates from the epidemic of post-surgical wound infections in Brazil have been sequenced (89, 91).

The epidemic of post-surgical wound infections in Brazil began in the city of Belém in the northern state of Pará in 2004 (374). Over the following decade, over 2000 cases were reported in at least 15 states spanning geographically distant regions of Brazil, with the highest concentration of cases observed in Rio de Janeiro, where 1051 cases were reported between 2006-2007 (130, 373, 375-378).

Molecular typing techniques found that isolates associated with the outbreaks were caused by *M. a. massiliense.* Partial *rpoB* sequencing of outbreak isolates from different locations revealed identical sequences and furthermore pulsed phase gel electrophoresis (PFGE) patterns from *DraI* digested DNA of isolates from different outbreak locations revealed identical patterns apart from a ~50kb band (130). Both results suggested that a single clone was responsible for the outbreaks in geographically distant regions in Brazil and that this clone may have adapted to thrive in this specific niche (130, 375).

The WGS representatives of the epidemic lineage from Goiás, GO-06, and Rio de Janeiro, CRM-019 and CRM-020, were found to be closely related to each other and to isolates from the pulmonary infection outbreak that occurred in Papworth hospital in the UK (89, 91). However, the shorter genetic distance between the outbreak isolates from Rio de Janeiro and Papworth (reported as one SNP in (89)) in comparison to the genetic distance between the isolates from Rio de Janeiro and Goiás (reported as 75 SNPs) suggested that there may

be more diversity between the outbreak isolates from Brazil than originally thought and that potentially a single lineage was not responsible for outbreaks (89).

On the other hand, epidemiological similarities were evident between the post-surgical wound infection outbreaks in Brazil. The vast majority of infections occurred after video-assisted surgeries and specifically after surgeries in which the surgical tools had been disinfected through immersion of the equipment for 30 minutes in 2% glutaraldehyde (GTA) solution (130, 373, 375-378). The epidemic lineage was found to be tolerant to GTA solutions up to concentrations of 8% (131). Observations of poor practices, including inadequate cleaning and sterilization of surgical equipment and the re-use of disposable equipment in some hospitals where outbreaks occurred led to the hypothesis that the epidemic lineage may have been exposed to nonlethal levels of GTA resulting in it adapting to become GTA tolerant and giving the lineage a selective advantage over other NTM lineages in the surgical environment (374). Furthermore, the epidemic lineage was also found to be more virulent than the *M. a. massiliense* type strain CIP108297 (379).

The GTA tolerance and increased virulence phenotypes of this lineage have yet to be fully explained, although potential candidates have been identified. A 56,264bp incP-1β circular plasmid, pMAB01, was identified in an isolate, INCQS 00594, from the initial outbreak in Belém (87). pMAB01 was predicted to encode 64 CDSs, with two variable regions encoding a mercury resistance transposon and streptomycin, kanamycin and sulfonamide resistance genes as well as a qacEdelta1 gene, which potentially encodes a protein that can export a range of drugs and biocides (87). Whist pMAB01 might be associated with the increased virulence of this lineage, it has been shown not to be responsible for the GTA tolerance phenotype of the epidemic lineage. In their investigation into the GTA tolerance phenotype, Lorena et al. used outbreak isolates with pMAB01 both present and absent and still observed the GTA tolerance phenotype in all the outbreak isolates (131). Consequently, this phenotype remains unexplained. Defects in porins have been suggested as a possible cause as GTA is believed to act by forming cross-links with proteins on the cell membrane and defects in porins were found to be responsible for the GTA tolerance of *M. chelonae* (131, *287)*. Genes with known virulence properties in other Mycobacteria, for example MmpL proteins, were found to have accumulated 86 unique SNPs in the outbreak isolate CRM-0020 in comparison to the Papworth and Seattle outbreak isolates (91). These SNPs could potentially be associated with the adaption of this lineage to increased virulence as well as to its adaption to the specific niche it is flourishing in in Brazil.

It also remains unclear how the lineage has spread to multiple hospitals, in multiple states throughout Brazil because in all the outbreaks the source was not identified (130, 373, 375-378). Surgeons moving with their own surgical equipment between different hospitals, cities and states has been proposed as a possible vehicle for transmission as has the possibility that the lineage had been disseminated to the geographically distant locations through contamination of non-activated GTA solution and the possibility that the lineage spread to distant outbreak locations through aquatic environments (131, 373).

Therefore, the aim of this project was to use WGS to understand the spread of the epidemic lineage throughout Brazil and to use comparative genomic techniques to determine how the epidemic lineage has adapted to the surgical niche, specifically its adaption to GTA tolerance.

5.2 Materials and Methods

5.2.1 Collection, DNA extraction and sequencing of post-surgical wound infection isolates from Brazil

Dr Sylvia Leão orchestrated the collection of 190 samples associated with the epidemic of post-surgical wound infections in Brazil, with the samples cultured from swabs taken from skin lesions or biopsies. The samples were collected from 9 different states between 2004 and 2010 (Appendix table 4.1). Vinicius calado Nogueira de Moura extracted the DNA from this collection using the following method: The isolates were grown in liquid Middlebrook 7H9 medium supplemented with OADC (oleic acid, alumin, dextrose and catalase) (Becton Dickinson). DNA from single colonies was extracted using the QIAamp DNA mini kit (Quiagen) according to the manufacturer's recommendations.

DNA sequencing was carried out using the Illumina Hiseq 2500 platform. Illumina libraries were constructed with a 450bp insert size according to Illumina protocols and used to generate 125bp paired-end sequences. One isolate was found to be contaminated and one isolate failed the initial round of sequencing. Both sequences were subsequently discounted from further analysis. In total 188 isolates associated with the epidemic of post-surgical wound infections in Brazil were analysed.

5.2.2 Mapping, de novo assembly and annotation

The reads of the 188 isolates from the epidemic of post-surgical wound infections in Brazil and the 526 isolates sequenced for Bryant et al's (2016) global population study were

mapped to the *M. a. abscessus* type strain ATCC19977, using BWA-MEM (73, 215). Variants were called using Samtools (v1.2) and Bcftools (v1.2) with the parameters described in chapter 7.3 (216).

To gain a higher resolution understanding of the genetic diversity between the Brazilian epidemic isolates and their closest relatives in the global population, including publicly available isolates previously sequenced from Brazil, CRM-0020 and GO-06, and the UK, 47J26, a reference was selected from the newly sequenced isolates. An isolate sampled from the initial outbreak in Belém, Pará, BRA_PA_42, was selected as the reference because the *de novo* assembly consisted of two contigs, with the chromosome present in a single contig. In total, the reads of 246 (Brazil: 188, publicly available: 3, closest relatives in global population: 55) isolates, were mapped, using BWA-MEM, to BRA_PA_42, with variants called via Samtools (v1.2) and Bcftools (v1.2) with the parameters described in section 7.3 and 7.4 (215).

Draft genomes were generated and annotated for the 188 isolates sequenced from the Brazilian epidemic following the methods described in section 7.6 and section 7.7 (219). Where necessary the Prokka annotations of the genomes were supplemented by comparing the gene sequences against the Pfam, InterPro and PHAST databases (220, 224, 225, 227).

5.2.3 Phylogenetic analysis

From the alignment produced after mapping the 526 isolates representative of the MABSC global population and the 188 isolates from Brazil (n=714) to the *M. a abscessus* ATCC19977 reference genome. 326,792 variable positions were identified with SNP-sites (217). RAxML (v.8.2.8) was used to infer a maximum likelihood phylogenetic tree from the alignment of these variable positions, as described in section 7.5 (61).

1,127 variable positions were identified with SNP-sites from the alignment produced after mapping the Brazilian isolates and the 58 isolates from the two clades in the global population most closely related to them to the Brazilian reference genome BRA_PA_42 using the methods described in section 7.3 and 7.4 (217). A maximum likelihood phylogeny with 100 bootstrap replicates was inferred via RAxML (v.8.2.8) from these variable positions (61).

To examine the diversity between the Brazilian epidemic isolates Minimum Spanning Trees (MST) were inferred using the goeBURST algorithm implemented in Phyloviz. This required

an alignment of the variable positions between the 188 Brazilian isolates when mapped to BRA_PA_42. This alignment was produced using BCFtools (v1.2) with variants called when the following parameters were met: a base call quality of greater than or equal to 10 and where at least four high quality reads supported alternative alleles (216). 197 variant sites were identified and assigned to 97 genotypes using the adegenet and poppr R packages (380, 381).

5.2.4 Temporal analysis

To determine whether there was a molecular clock within the Brazilian epidemic lineage, the phylogenetic root-to-tip distance was regressed against the sampling date, with the correlation examined with the phylogeny rooted to maximize r^2, maximize signed r^2 and minimize residual mean squares (382). Isolation dates were permuted 1000 times using a clustered permutation approach to account for the potential of confounding temporal and genetic structures (i.e the likelihood that closely related sequences have been sequenced at the same time) to obtain the statistical significance of the regression (382). This required a phylogeny inferred from the variable positions present in a recombination free alignment of the 173 Brazilian isolates where the day/month/year of isolation was available. This alignment was produced using Gubbins (383). The 144 variable positions present in the 4,686,049 bp recombination free alignment were extracted using SNP-sites (217). The phylogeny was inferred using RAxML (v.8.2.8) (61).

Bayesian Markov Chain Monte Carlo (MCMC) analysis, implemented using the BEAST (v.1.8.3) package, was used to estimate the date of emergence of the Brazilian epidemic lineage (384). BEAUTi was used to create the XML file required by BEAST from the recombination free SNP alignment of 173 Brazilian isolates (384). Three independent MCMC chains of 100 million states were run using a GAMMA site heterogeneity model with a relaxed log normal clock and constant population size model. The convergence after an initial burn-in period of 10%, the agreement between the three runs and determination of whether all effective sample size (ESS) values were greater than 200 were assessed using Tracer (v1.6) (385).

5.2.5 Comparative genomic analyses

Assemblies consisting of less than 100 contigs and with lengths that fell within 1.5 times the interquartile range were used to compare the genome sizes of the Brazilian lineage to the size of the three MABSC subspecies. Large scale insertion and deletions (indels) between the Brazilian lineage reference, BRA_PA_42, and representatives of the two clades most

closely related to the lineage in the global population were detected using blastn (v2.4.0) comparisons (386). The results were visualized using the artemis comparison tool (ACT) (297). The Prokka annotations of the gene content of the indels was enhanced by comparing the gene sequences with the Pfam (v.3.1.0), InterPro (v.68) and PHAST databases and blastn comparisons against the non-redundant nucleotide sequence database (224, 225, 227). GO-terms were assigned using BLAST2GO (233).

5.2.6 Analysis of the second contig in the new Brazilian reference genome BRA_PA_42

The Brazilian lineage reference genome selected from the newly sequenced isolates, BRA_PA_42, consisted of two contigs, with the chromosome present in a single contig. The second contig was compared using blastn to the non-redundant nucleotide sequence database. To test the hypothesis that this second contig was a novel plasmid, the potential circularity of this contig was examined by looking at read pair information, whilst the prokka annotation of the contig, supplemented by comparing the gene sequences against the Pfam (v.3.1.0), InterPro (v.68) and PHAST databases, was examined for core plasmid genes and a lack of core phage genes.

The gene order of the Type VII secretion system (T7SS) encoded by the second contig was compared to the T7SSs recently described by Ummels et al. (2014) and Dumas et al. (2016) and to the chromosomal ESX-5 T7SS encoded by *Mycobacterium tuberculosis* H37Rv (387, 388). The average nucleotide identity (ANI) and average amino acid identity (AAI) between these T7SSs was calculated using blastn and tblastx, with the following parameters: an E-value of less than or equal to 0.00001 and match length greater than or equal to 100.

The prevalence of both this second contig and the plasmid previously found to be associated with the Brazilian post-surgical wound infection epidemic, pMAB01, in the MABSC global population was determined by mapping the reads of the 714 isolates (global population: 526, Brazil: 188) to this contig and the reference of pMAB01 (CP003376.1) with BWA-MEM (87, 215). The presence or absence of the plasmid across all the isolates used in this study was also determined by comparing the assemblies to a local blastn database built from the nucleotide sequences of the second contig in BRA_PA_42 and pMAB01.

5.2.7. Investigating mycobacterial porins as the genetic basis of the GTA resistant phenotype displayed by the Brazilian lineage

5.2.7.1 Identifying msp porins in M. a. abscessus ATCC19977

Mycobacterium smegmatis porins (msp*)* have been shown to be responsible for resistance to aldehyde based disinfectants in *M. chelonae* and *M. smegmatis* (287). To test the hypothesis that these porins could be playing a role in the GTA resistant phenotype of the Brazilian lineage a tblastx comparison was carried out, and visualized using ACT, between *M. a. abscessus* ATCC19977 and *M. chelonae* ATCC 35752 to identify the porin region corresponding to the *mspA-C* loci in *M. chelonae* (MCH_4689c, MCH_4690c and MCH_4691c) that has been associated previously with resistance to aldehyde disinfectants (287). The CDSs MAB_1080 and MAB_1081 were found to have high AAI (96%) with the three *M. chelonae* porins and occurred in the same wider sequence context.

5.2.7.2 Investigating the genome organisation at the mspA/mspB porin locus

To investigate the genome organization at the locus encoding *mspA* (MAB_1080) and *mspB* (MAB_1081) across the MABSC global population, the reads of the 715[11] isolates that make up the MABSC global population dataset including the Brazilian epidemic lineage were mapped, using BWA-MEM, to the 3,803 bp region encoding the *mspA* and *mspB* porin genes in *M. a. abscessus* ATCC19977 (215). The coverage over the porin locus containing the *mspA* and *msp*B genes was determined for each isolate using Samtools v1.3, with a mapping quality threshold of 20 (probability mapped to correct location 99%) and base call quality of 30 (base call is 99.9% accurate) (216).

Presence of both *msp*-like porin genes at this locus was determined by an average depth of coverage greater than eight in the 436bp intergenic region between *mspA* and *mspB* and an average depth of coverage greater than eight over both *mspA* and *mspB*. Deletion of both *msp*-like porin genes at this locus was determined by the absence of coverage (average of less than eight reads mapped) in the intergenic region between *mspA* and *mspB* and a reduction in depth of coverage greater than 20 between the 1000bp region upstream of the start of the porin loci and the 94bp signal peptide region of *mspA*. The presence of a single *msp*-like porin gene at this locus was determined by absence of coverage in the intergenic region between *mspA* and *mspB* (average of less than eight reads mapped) and no

[11] The Brazilian isolate that failed the initial round of sequencing was re-sequenced and subsequently included in the analysis of the porin genes in the Brazilian epidemic lineage

reduction in depth of coverage greater than 20 between the 1000bp flanking region upstream of the porin loci and the 94bp signal peptide region of *mspA*.

5.2.7.3 Generation of porin knock-out mutants

The following work was carried out by Vinicius calado Nogueira de Moura. The *Mycobacterium abscessus* subspecies *massiliense* type strain CIP108297 was grown at 37℃ in Middlebrook 7H9-OADC broth (BD, Difco) supplemented with 0.005% Tween 80, Mueller-Hinton II broth (BD, Difco), minimal Glycerol-Alanine-Salt (GAS) medium supplemented with 0.05% tyloxapol (pH 6.6), or on Middlebrook 7H11-OADC agar (BD, Difco). Kanamycin (Kan), hygromycin (Hyg) and streptomycin (Str) were added to final concentrations of 400, 1,000 and 200 µg/ml respectively. The strain used for cloning was *Eschericha coli* DH5α. This was grown in LB Lennox (BD, Difco) medium at 37℃.

The orthologs of *mspA* and *mspB* in *M. a. massiliense* CIP108297, MMCCUG48898_0905 (*mspA*) and MMCCUG48898_0906 (*mspB*) were inactivated using allelic replacement via a recombineering system. The mycobacteriophage Che9c recombineering proteins, Gp60 and Gp61, were expressed, under the control of an acetamide-inducible promoter, from the replicative plasmid pJV53-XyLE (316, 389). *M. a. massiliense* CIP108297 colonies harboring the acetamide-induced plasmid pJV53-XyIE were electro-transformed with linear allelic exchange substrates encoding the *mspA* and *mspB* loci. Double-crossover mutants were isolated on Str-containing medium. To delete the *mspA* locus, a linear allelic substrate was generated by bookending the streptomycin-resistance cassette from pHP45Ω with 571 bp of the DNA upstream of the internal XmnI restriction site of mspA and 500 bps of the DNA downstream of this site. Similarly, to delete the *mspB* locus, a linear allelic substrate was generated by bookending the streptomycin-resistance cassette with 571 bp of the DNA upstream of the internal XmnI restriction site of *mspB* and 500 bps of the DNA downstream of this site. To delete both *mspA* and *mspB*, the linear allelic substrate consisted of the streptomycin cassette bookended by the 571bp of DNA upstream of the XmnI site of *mspA* and the 357bp of DNA downstream of the XmnI site of *mspB*.

To complement the knock-out mutants the replicative plasmids pOMK-*mspA*, pOMK-*mspB* and pOMK-*mspAB* were constructed by cloning *mspA*, *mspB* or the 2.1 kb region encoding both *mspA* and *mspB* into the multicopy plasmid pOMK (389). The porin genes are expressed by their own promoter in these complementation constructs.

5.2.7.4 Glutaraldehyde tolerance assay

The susceptibility of the *M. a. massiliense* CIP108297 wild-type, *mpsA* knock-out, *mspB* knock-out, *mspAB* knock-out and their respective complemented isolates, to commercial GTA was determined by performing suspension tests following the protocol used by de Moura et al. (389). This work was performed by Vinicius calado Nogueira de Moura

5.3 Results

5.3.1 A single, recently emerged lineage is responsible for post-surgical wound infection outbreaks throughout Brazil

The Brazilian isolates formed a single clade, to the exclusion of all other MABSC isolates, within the global population phylogeny (Figure 48A), confirming that the epidemic was caused by a single lineage (Herein referred to as the Brazilian lineage). The clade was most closely related to the recently DCC in *M. a. massiliense* subspecies which encompasses the causal lineages of the Papworth and Seattle outbreaks (73) agreeing with previous descriptions of the position of the Brazilian lineage in the MABSC global population (89, 91).

The higher resolution phylogeny, inferred from the variant positions identified after mapping the Brazilian lineage, the isolates that made up the DCC and the outlying clade to BRA_PA_42, showed that there were 149 SNPs between the last common ancestor of the Brazilian lineage and its shared common ancestor with the DCC (Figure 48B). It also showed that there was evidence of geographical and temporal structure within the Brazilian epidemic lineage (Figure 48B).

Due to the temporal structure evident in the phylogeny shown in Figure 48B, root-to-tip linear regression analysis was performed to examine whether there was a temporal signal within the Brazilian lineage (382). A significant temporal signal was detected, using a phylogenetic root satisfying three separate criteria, within the Brazilian epidemic lineage (Figure 49) and therefore a more accurate estimation of the substitution rate and date of emergence of the lineage was determined using BEAST (384). This estimated that the Brazilian lineage emerged in 2003 (2002.7; upper 95%: 2003.8, lower 95%: 2001.3), just prior to the initial outbreak in Pará in 2004 (Figure 50) (374). The mean substitution rate was estimated to be 8.25×10^{-07} SNPs/site/year (upper 95%: 1.07×10^{-06}, lower 95%: 5.31×10^{-07}) which equates to approximately 3.8 SNPs/genome/year (2.4–5.0 SNPs/genome/year), similar to that

previously measured for a DCC within *M. a. abscessus* (3.5×10^{-7} SNPs/site/year ($1.83\text{x}10^{-5}$,11×10^{-7})) (73).

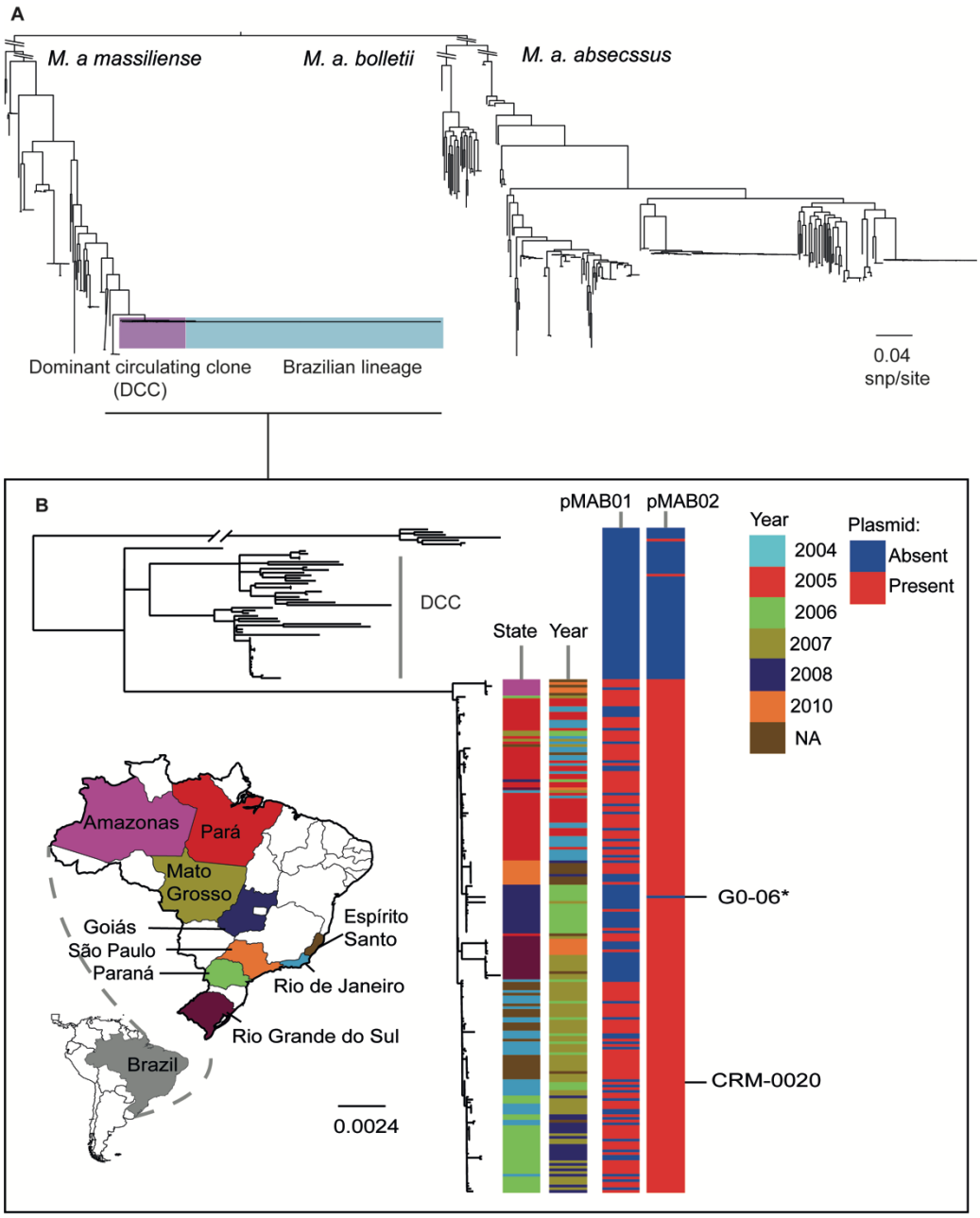


Figure 48: A single lineage of *M. a. massiliense* is responsible for the epidemic of postsurgical wound infections in Brazil

A) Midpoint rooted maximum likelihood phylogeny of the Brazilian epidemic lineage in the context of the *Mycobacterium abscessus* species complex global population. The Brazilian isolates form a single clade (light blue), closely related to the *M. a. massiliense* DCC.
B) Maximum likelihood phylogeny of the Brazilian outbreak lineage, the closely related DCC rooted to an outlying clade. The metadata columns show that the Brazilian outbreak lineage isolates cluster by state and year and show the presence (red) and absence (blue) of the two plasmids associated with this lineage, pMAB01 and pMAB02.

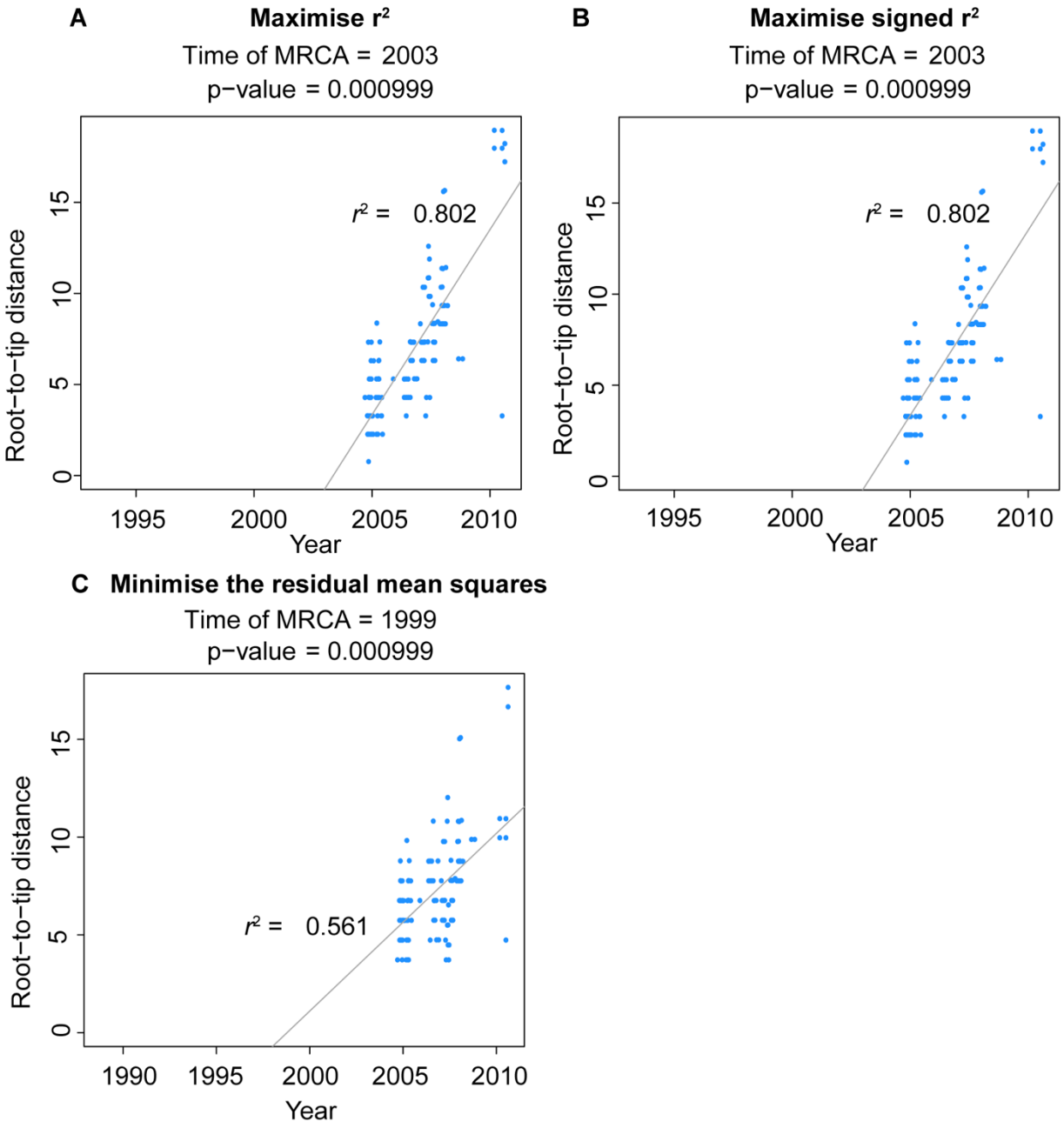


Figure 49: Significant temporal signal detected within the Brazilian epidemic lineage

Linear regression plots of the phylogenetic root-to-tip distance against sampling dates. The dates were randomly permuted, using clustered permutation, 1000 times to determine the significance of the correlation. Each plot shows the correlation between the root-to-tip distance and the sampling date when the phylogeny was rooted A) to maximise r^2 B) to maximise signed r^2 and C) to minimise the residual mean squares of the model.

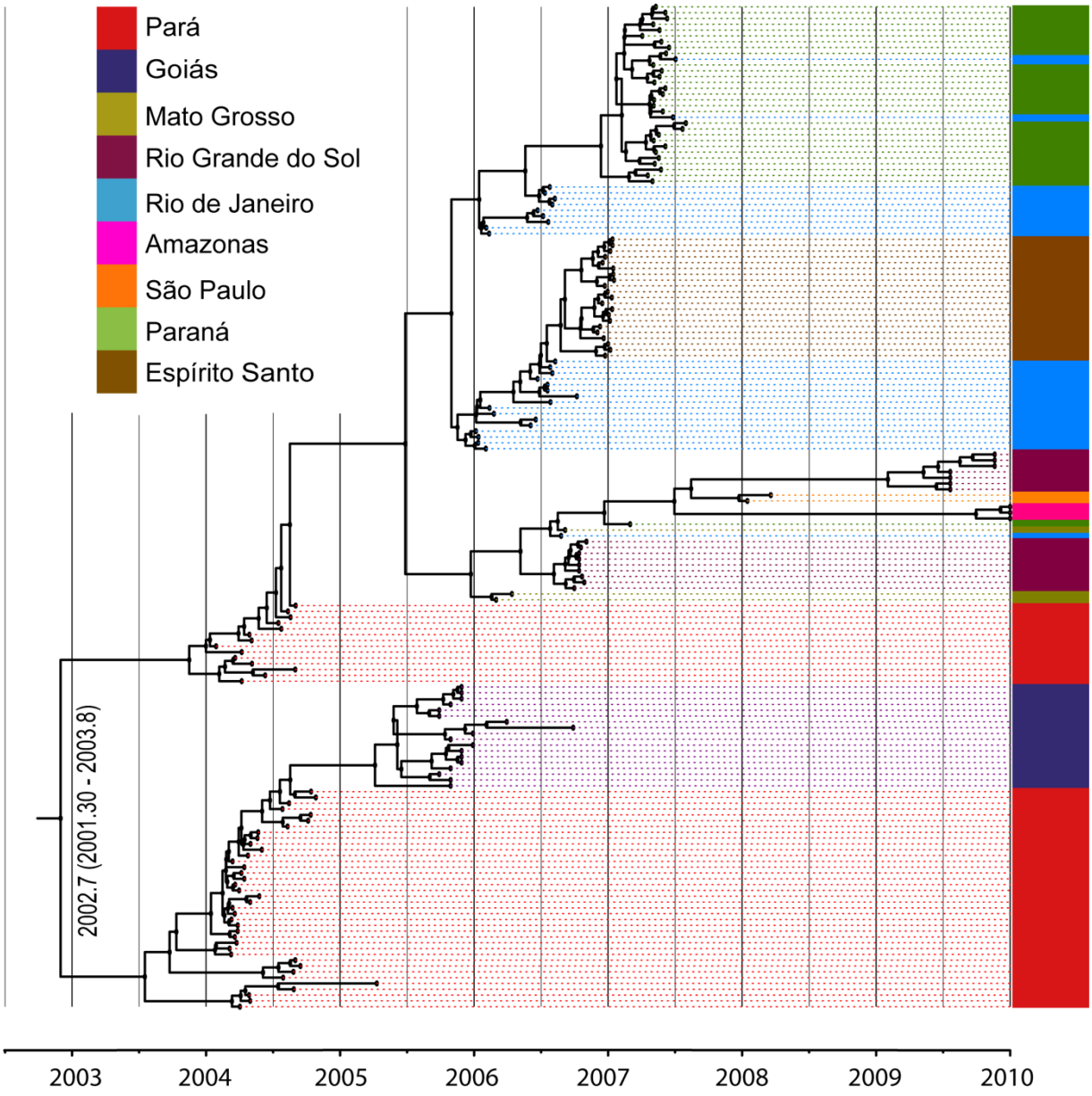


Figure 50: Brazilian lineage emerged in 2002, just prior to the initial outbreak in Belem

Maximum clade credibility tree of the Brazilian lineage inferred from the 27,003 trees produced by BEAST. The emergence of lineage was estimated by BEAST to be approximately 2003. The colors represent the state from which the isolate was collected.

5.3.2 Onward spread throughout Brazil

Geographical clustering was also evident within the phylogeny shown in Figure 48B. As this suggested the Brazilian epidemic of post-surgical wound infections was not caused by a point source outbreak MSTs were constructed to determine the potential transmission route of the Brazilian lineage between different states within Brazil. The MSTs, shown in Figure 51,

show that the state in which the original outbreak occurred, Pará, acts as the main hub for transmission throughout the epidemic period. Transmission of the lineage from Pará potentially seeded the subsequent outbreaks in Goiás (2006), Mato Grosso (2006), Rio Grande do Sul (2007) and Amazonas (2010). There is also evidence that a transmission event occurred between Pará and Paraná, however, the majority of cases in Paraná appear to have been seeded by a transmission event from Rio de Janeiro. Whilst a transmission event from Pará in all likelihood introduced the lineage into Espírito santo (2006-2007) and/or Rio de Janeiro (2006-2007) it is not possible to determine which location it was introduced into first. Similarly, it is unclear whether a transmission event from Pará or Goiás seeded the outbreak in São Paulo.

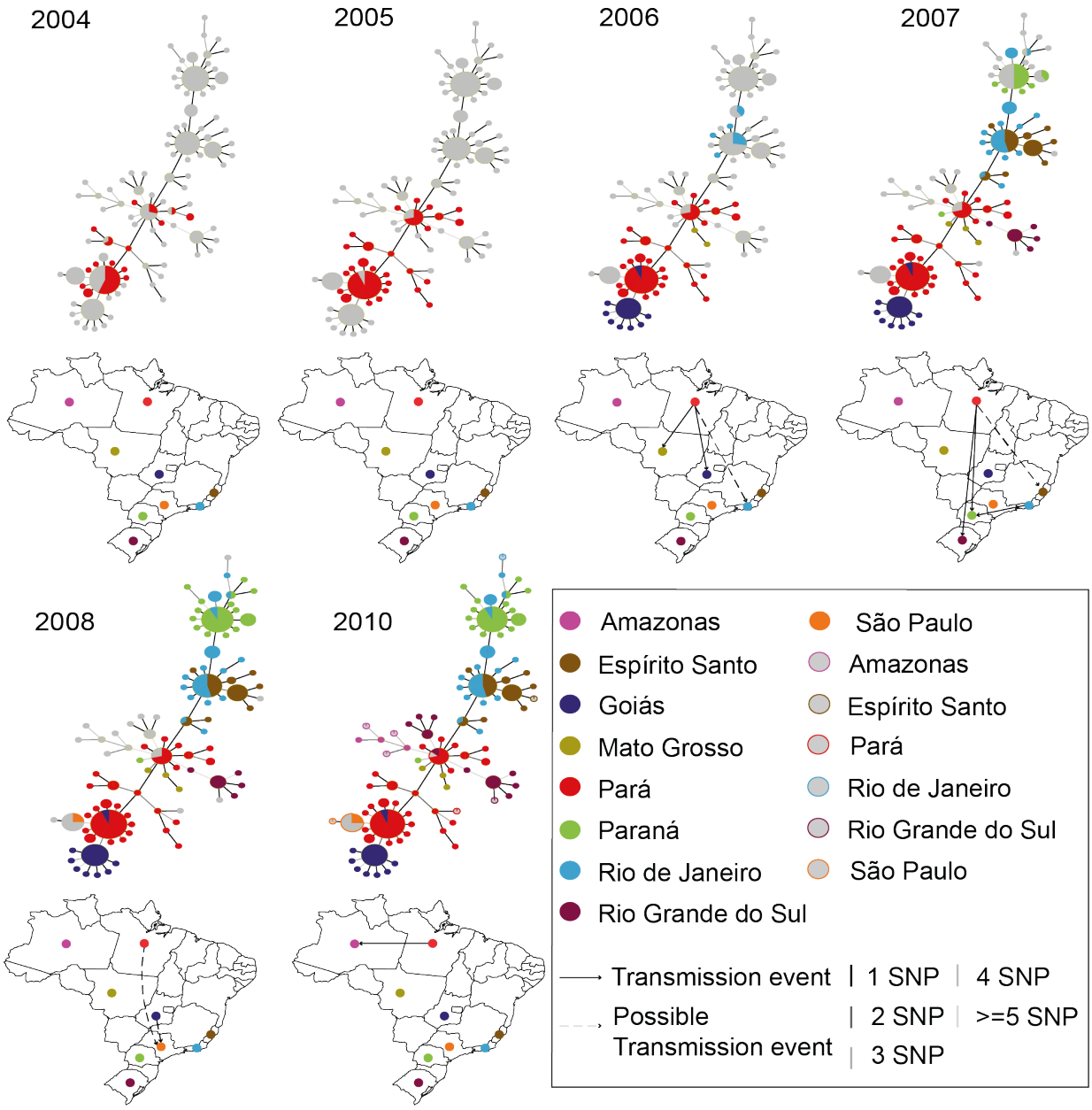


Figure 51: Brazilian epidemic lineage spread to geographical distant parts of Brazil *via* several waves of transmission out of Pará

Minimum spanning trees (MSTs) showing the transmission route supported by the genetic data. The node sizes represent the number of isolates. The nodes are colored according to the state from which the isolate was collected. Open circles represent isolates with a known location, but the date of collection of the isolate was unknown. The MSTs imply that transmission events from Pará introduced the epidemic lineage to Goiás, Mato Grosso and either Rio de Janeiro or Espirito Santo in 2006, Rio Grande do Sul and Paraná in 2007 and Amazonas in 2010. The outbreak in São Paulo in 2008 was seeded either from Goiás or Pará. A transmission event from Rio de Janeiro to Paraná in 2007 is shown to be responsible for the majority of cases in this region.

5.3.3 Large scale deletions characterize the Brazilian epidemic lineage and suggest it's undergone adaptation to a novel niche

The genetic data presented in this study and epidemiological investigations presented in previous studies both suggest that a single lineage has spread to multiple different states in Brazil, which implies that this lineage may have adapted to gain a selective advantage in the Brazilian surgical niche (130). This in part has been proved by the identification of the GTA resistance phenotype displayed by this lineage (131). Blastn analysis between representatives of the Brazilian lineage, the DCC and the outlying clade showed that Brazilian lineage had incurred 14 deletions (Figure 52), however only two of these deletions were unique to the Brazilian lineage. This led to the hypothesis that the Brazilian lineage had potentially undergone reductive evolution and therefore the genome sizes of the Brazilian lineage were compared to the genome sizes of the isolates that make up the rest of the global population grouped by subspecies. Figure 53 shows that the Brazilian lineage was on average 298,431 bps smaller than the average genome size of the three subspecies. As the smaller genome size was potentially indicative of reductive evolution and could represent the adaption of the lineage to a novel niche, the CDS content of the two deletions that had only occurred in the Brazilian lineage was investigated to determine the potential functions lost by the lineage. The deletion, del_12078_1#71_01025_01041, consisted of 17 CDSs, with a functional annotation, reported in appendix table 4.2, possible to predict for seven. The deletion, del_12078_1#71_02602_02653, consisted of 52 CDS, with the functional annotation predicted for 49, reported in appendix table 4.3. Neither deletions were found to encode CDSs with significant nucleotide similarity to known phage genes in the PHAST database. However, del_12078_1#71_02602_02653 shared 89% ANI to the previously described insertion sequence ISMab1 (85).

GO-term analysis showed that del_12078_1#71_02602_02653 encoded 13 CDSs involved in oxidation and reduction processes and that nine CDSs were associated with metabolic processes. Three of the 17 CDSs encoded by del_12078_1#71_010205_01041 had GO-term functions suggesting they played in role in DNA integration processes, whilst seven were predicted to be involved in DNA-binding processes. The possible function of the remaining six CDSs present in this deletion could not be predicted.

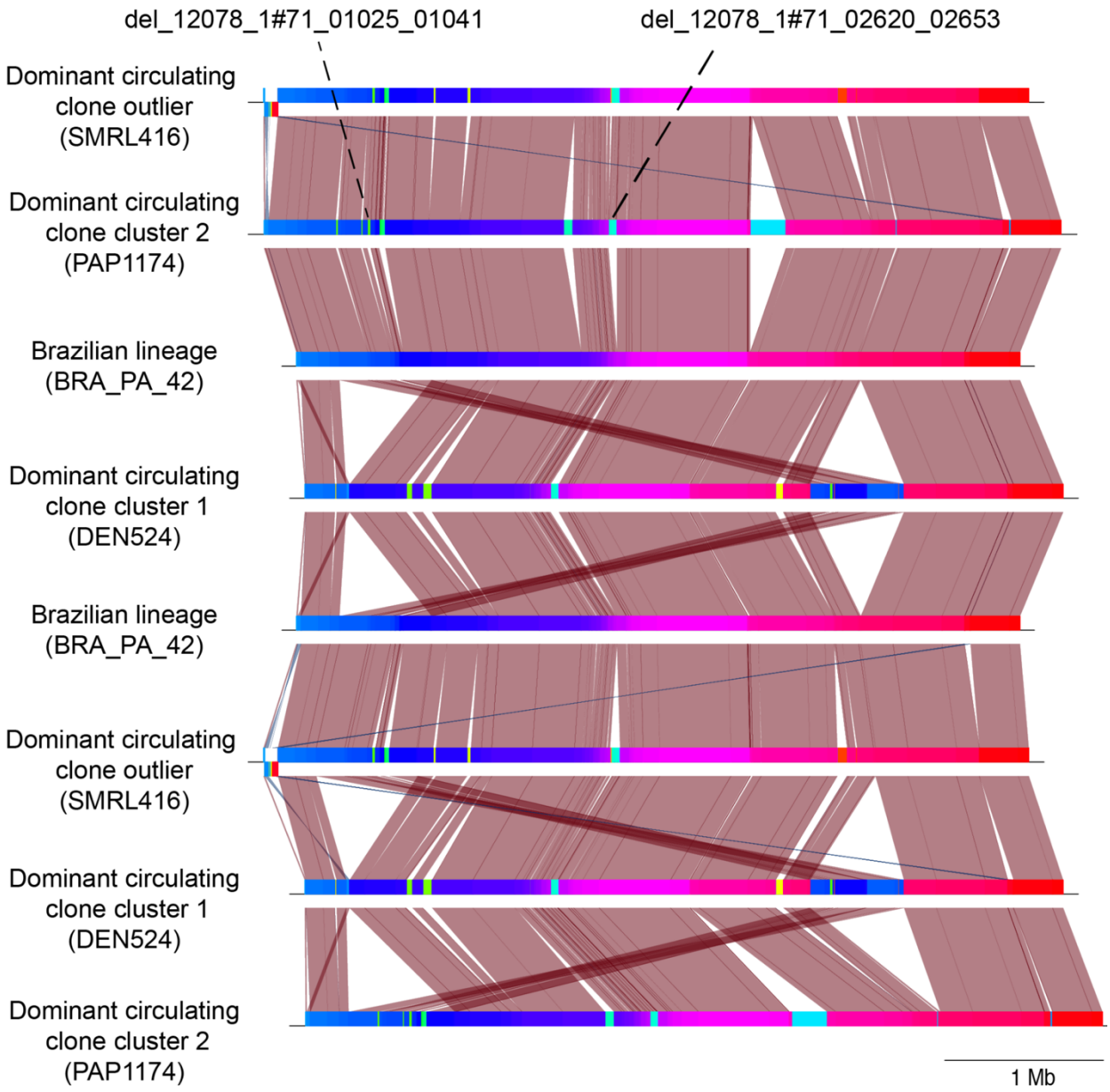


Figure 52: Large scale deletions characterize the Brazilian lineage

Blastn comparisons between the Brazilian lineage reference genome, BRA_PA_42 and two representatives of the closely related dominant circulating clone and a representative of an outlying clade. 14 deletions were detected to have occurred in the Brazilian lineage, but only two, labelled, had been deleted solely in the Brazilian lineage.

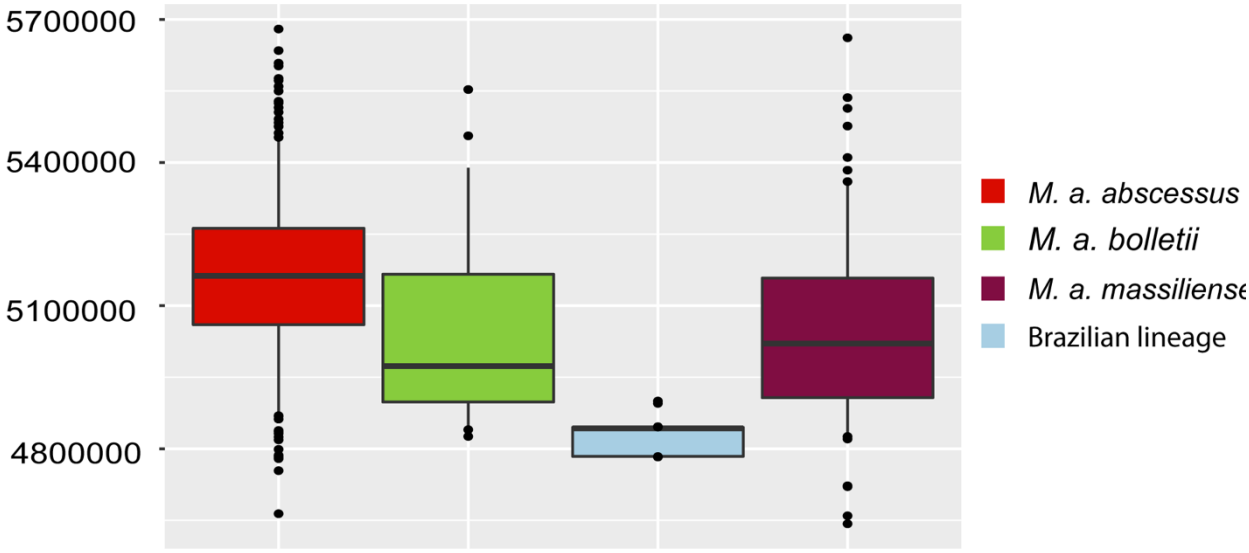


Figure 53: Brazilian epidemic lineage has a smaller genome size than the average for the MABSC

Boxplots comparing the genome sizes of the three subspecies (*M. a. abscessus* (red): 352, *M. a. massiliense* (purple): 130, *M. a. bolletii* (green): 30) to those of the Brazilian lineage (blue). The borders of the boxplots represent the 25th and 75th percentile, the bold line represents the median genome size and the whiskers mark the 5th and 95th percentile. The median size of the Brazilian lineage falls between the 5th and 25th percentile for *M. a. massiliense* and below the 5th percentile for *M. a. abscessus* and *M. a. bolletii*, suggesting it is markedly smaller than the majority of MABSC lineages.

5.3.4 A second novel plasmid, pMAB02, is associated with the Brazilian epidemic lineage

There was also evidence that the Brazilian epidemic lineage had adapted to a novel niche through the gain of genetic material, with the presence of the previously described incP-1β plasmid, pMAB01, and a potentially novel 95,804 kb plasmid, identified in BRA_PA_42 (87). The nucleotide sequence of this potentially novel plasmid, the second contig in the reference genome BRA_PA_42, when compared against the non-redundant nucleotide sequence database was found to share 99% nucleotide identity with a query coverage of 100% to a sequence described as *Mycobacterium* phage *Adler* (KC960489.1). However, the presence of core plasmid genes such as *parA*, *parB*, *traI* and *traM* and a lack of core phage genes (appendix table 4.4) as well as evidence from read pair information that it was circular suggested that this contig was most likely a plasmid, which was subsequently designated pMAB02.

The plasmids, pMAB01 and pMAB02, were found to differ in their presence both within the Brazilian epidemic lineage and across the global population. pMAB01 was present in 62% (117/190[12]) of the Brazilian isolates and no further isolates in the global population (Figure 48B, Figure 54A). Figure 48B showed that there was potentially evidence of geographical loss of pMAB01, with a sub-clade consisting of only isolates from Rio Grande do Sul all having lost the plasmid. pMAB02 was present in 99% of the Brazilian isolates and only five other isolates in the global population (Figure 48B, Figure 54A). GO-06, marked in Figure 48B, was the only isolate from Brazil found not to contain pMAB02, however, this isolate was re-sequenced in this study (BRA_GO_06) and the plasmid was found to be present.

The presence of pMAB02 in all the isolates from Brazil suggested it could be conferring a selective advantage upon the lineage, therefore the gene content of the plasmid was investigated for potential virulence determinants. pMAB02 encoded 121 CDSs (Figure 54B) but it was not possible to predict the function of 89 of the CDSs. However, included amongst the 32 CDS for which it was possible to predict a function were plasmid partitioning proteins, *parA* and *parB*, and further genes involved in conjugation (appendix table 4.4). Furthermore, a complete T7SS was encoded by pMAB02 (Figure 54B). Gene order comparisons between the pMAB02 T7SS and the T7SSs encoded by *M. a. bolletii* 50594 plasmid 2, *Mycobacterium marinum* plasmid pRAW and the chromosomally encoded ESX-5 T7SS found on *M. tuberculosis* H37Rv showed that the pMAB02 shared most similarity to the recently described ESX-P1 T7SS, that had originally been identified on *M. marinum* plasmid pRAW (Figure 54C) (387). No nucleotide similarity (i.e no blastn hits were returned with an e-value less than the cutoff of 0.00001) was detected between the T7SS encoded by pMAB02 and these T7SSs. However, the average amino acid identity (AAI) between the pMAB02 T7SS and the T7SSs encoded by *M. a. bolletii* 50594 plasmid 2, *Mycobacterium marinum* plasmid pRAW and the chromosomally encoded ESX-5 T7SS was calculated to be 36%, 38% and 37% respectively.

[12] 188 isolates sequenced in this study and two publicly available isolates, GO-06 and CRM-0020

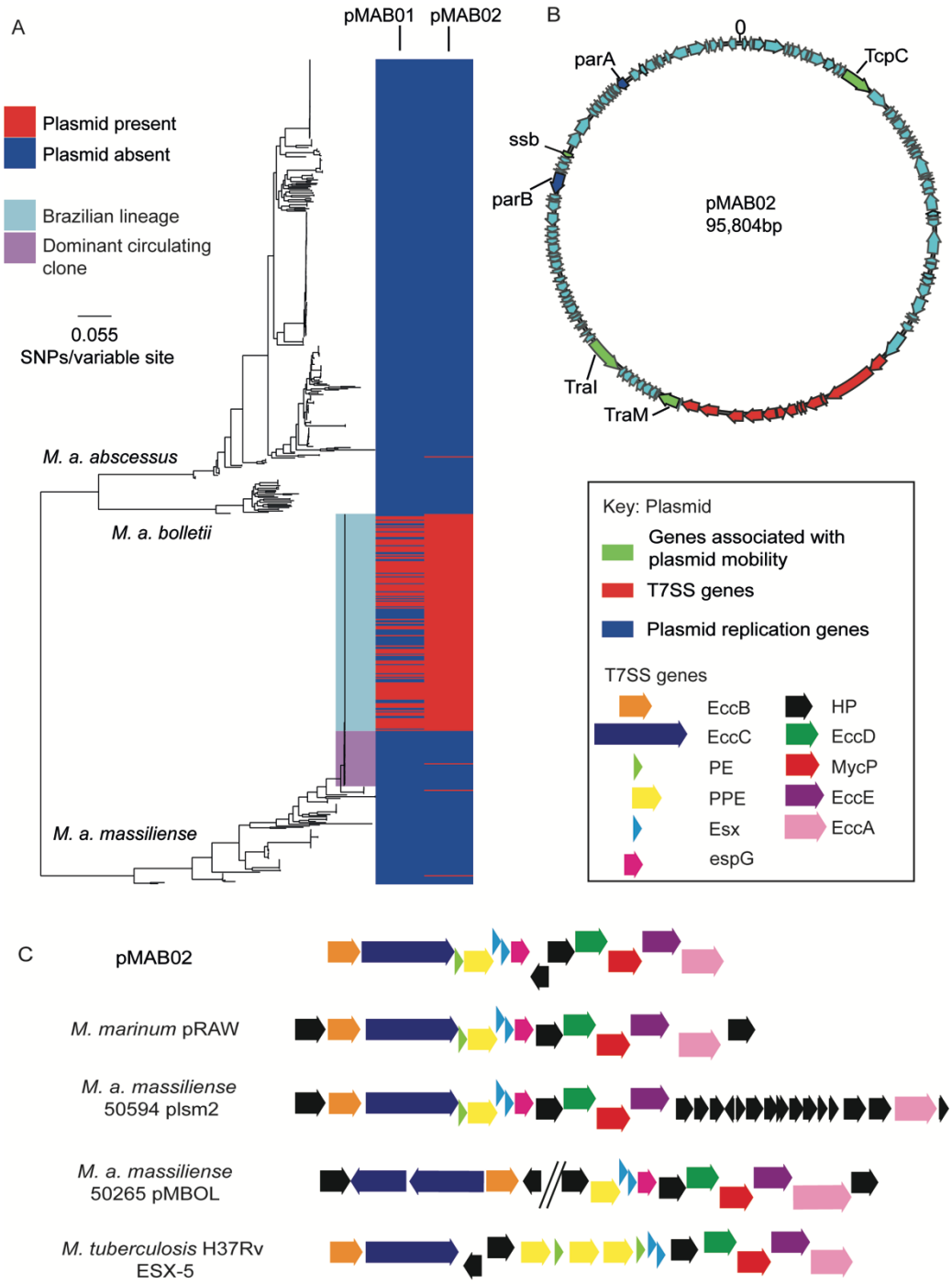


Figure 54: A second novel plasmid encoding a type VII secretion system present in the Brazilian epidemic lineage

A) Presence (red) and absence (blue) of the plasmids, pMAB01 and pMAB02, across the MABSC global population. B) Plasmid map of the newly described pMAB02. Genes associated with plasmid mobility (green) and replication (blue) and the T7SS (red) are highlighted. C) Gene order comparison between the T7SS encoded by pMAB02 and those encoded by either other Mycobacterial plasmids or other Mycobacterial chromosomes. The T7SS has the most similar gene order to the T7SS encoded by pRAW.

5.3.5 Variation in the genome organization at the mspA/mspB porin locus in the Brazilian epidemic lineage

Defects in porin genes have been associated with biocide resistance in other *Mycobacterium* species (287). Given that the GTA tolerance phenotype of the Brazilian epidemic lineage remains unexplained, with pMAB01 having been found not to be associated with this phenotype and no likely candidates on the novel plasmid pMAB02, the two *msp*-like porins, MAB_1080 and MAB_1081, that were shown through tblastx analysis to share 96% AAI with the *msp* porins associated with aldehyde based disinfectant resistance in *M. chelonae* and to be encoded in the same sequence context, were investigated (87).

The reference porin region (1093346bp to 1097148bp) extracted from *M. a. abscessus* ATCC19977 included the two *msp* orthologs and 1000bp flanking regions up and down stream. The resulting reference was 3,803bps in length. MAB_1080 consisted of 672bps and 223 amino acids whilst MAB_1081 consisted of 669bps and 222 amino acids.

Analysis of the depth of coverage of the mapped reads over the *mspA* and *mspB* porin locus showed that of the 526 isolates associated with pulmonary infections, 469 isolates encoded two *msp*-like porin genes at this locus (Figure 55, Figure 56A), 54 encoded a single fusion *msp*-like porin gene (Figure 55, Figure 56B) and three isolates encoded no *msp*-like porin genes at this locus (Figure 55, Figure 56C). A single *msp*-like porin gene, a fusion of *mspA* and *mspB*, was observed in 180 of 189 isolates associated with the post-surgical wound infection epidemic in Brazil, the remaining 9 isolates encoded two *msp*-like porin genes at this locus (Figure 55).

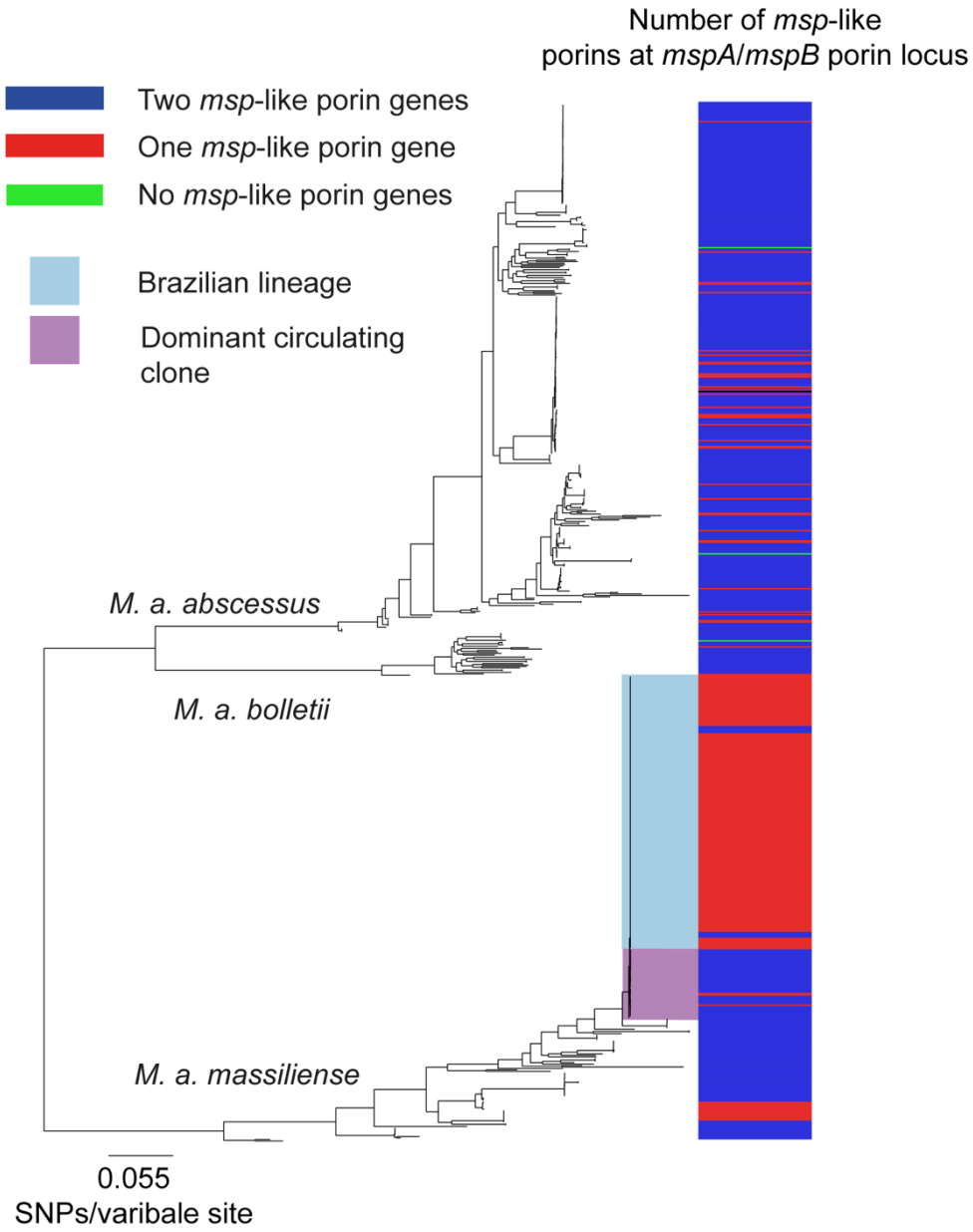


Figure 55: A fusion of *mspA and mspB* porin genes has occurred in the majority of isolates from Brazil

The maximum likelihood phylogenetic tree of the MABSC global population with the metadata column representing the number of *msp*-like porin genes encoded at the *mspA*/*mspB* porin locus. The majority of isolates in the MABSC global population encoded two *msp*-like porins at this locus. A deletion of a porin at the *mspA*/*mspB* porin locus has occurred in the majority of isolates from Brazil.

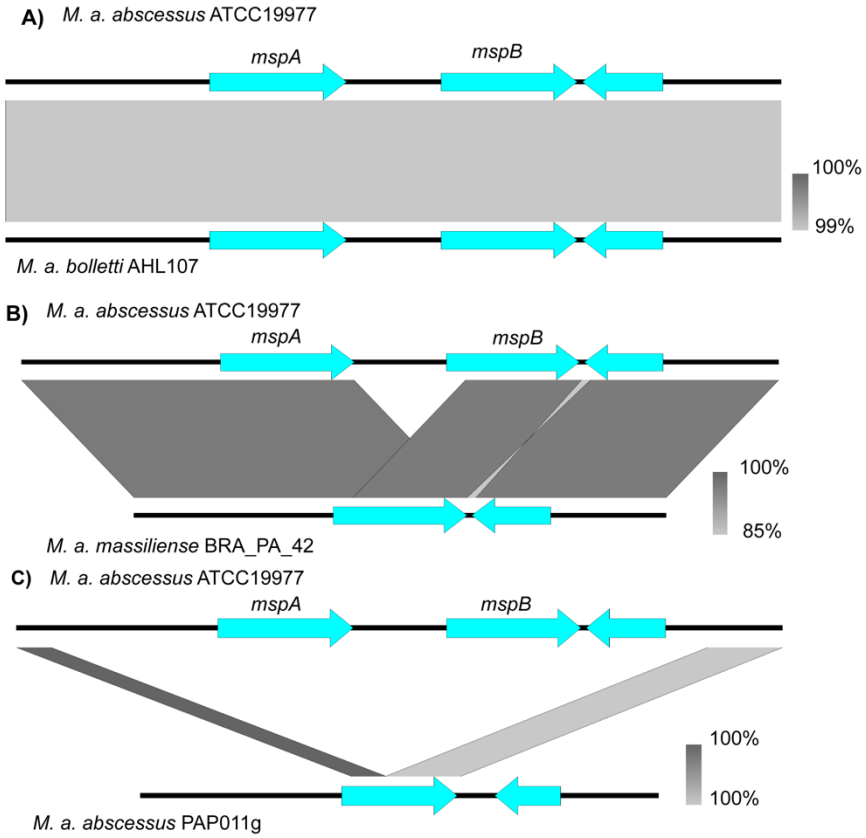


Figure 56: Variation observed across the MABSC global population at the mspA/mspB porin locus

Blastn comparisons, visualized with EasyFig (276), showing the genome variation seen across the MABSC global population at the *mspA*/*mspB* locus. A) this shows the genome organization at the locus when both *mspA* and *mspB* are present. B) This shows the genome organization at this locus after a recombination event has occurred resulting in the deletion of a porin. C) Represents an example of the deletion of both *mspA* and *mspB*.

5.3.6 Deletion of a porin gene at the mspA/mspB porin locus is not responsible for the GTA tolerance of the Brazilian epidemic lineage

Given that a porin gene was found to be deleted in 95% of the Brazilian epidemic isolates, *M. a. massiliense* CIP108297 porin knockout mutants and their respective complements were generated by Vinicius Calado Nogueira de Moura to test their susceptibility to GTA. All the knock-out mutants were found to be susceptible to GTA and to have similar GTA susceptibilities to that of their WT parent, with the complement mutants also presenting similar GTA susceptibilities (Figure 57).

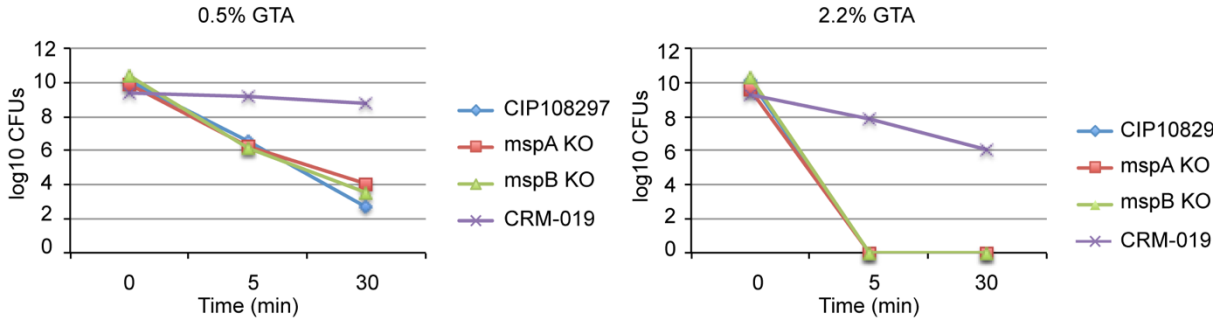


Figure 57: Deletion of a porin not responsible for the glutaraladehyde tolerance phenotype of the Brazilian epidemic lineage

Number of colony forming units (CFUs) after suspension of *M. a. massiliense* CIP108297 WT, *mspA* KO, *mspB* KO, and Brazilian lineage isolate CRM-0019 in 0.5% and 2.2% GTA solution after five and 30 minutes respectively. The CIP108297 porin KO mutants were found to be susceptible to killing by GTA and mirrored the response of the WT at both concentrations tested, whist the Brazilian lineage representative CRM-0019 was shown to be able to survive exposure to both concentrations tested. Figure courtesy of Vinicius Calado Nogueira de Moura.

5.4 Discussion

An epidemic of post-surgical wound infections occurred in Brazil between 2004-2011 (130, 373, 375-378). Molecular typing techniques suggested that a single *M. a. massiliense* clone was responsible for the post-surgical wound infection outbreaks that occurred in as many as 15 states (130, 373). The outbreaks were epidemiologically linked by their association with video-assisted surgeries where the instruments were disinfected with GTA and the epidemic clone was shown to survive 30 mins exposure to GTA concentrations of less than 8% (131). The clone was also shown to be more virulent than the *M. a. massiliense* type strain CIP108297 (379). However, the sources of the outbreaks failed to be identified and it was unclear whether the epidemic was caused by a point source or whether it had been transmitted to the outbreak locations throughout Brazil. The molecular mechanism driving the GTA resistance phenotype was also unknown as were the genetic adaptations potentially driving the increased virulence displayed by this clone, although some candidates had been identified (91). Furthermore, at the commencement of this study only three isolates from this epidemic had been WGS and the results had suggested there was greater diversity between the outbreak isolates from different locations than suggested by molecular typing techniques (89, 91).

WGS can be used to understand the transmission dynamics of outbreaks and uncover the genetic adaptations driving life threatening pathogens. This study aimed to harness this ability by WGS isolates from nine outbreak locations spanning the duration and geographical breadth of the post-surgical wound infection epidemic to determine how the epidemic lineage spread throughout Brazil and how it had adapted to be more virulent and specifically GTA resistant.

Phylogenetic analysis confirmed that a single lineage was responsible for the epidemic of post-surgical wound infections in Brazil, as had been suggested by molecular typing (Figure 48A) (130). It also supported the close phylogenetic relationship between the epidemic in Brazil and the pulmonary infection outbreaks that occurred in CF centers in the UK and US (89, 91), which have subsequently been found to form part of a virulent and human transmissible *M. a. massiliense* lineage (DCC3) that has spread globally (73). The close relationship of the epidemic lineage to the DCC and the recent emergence of the lineage, estimated by BEAST to be c. 2003, suggests that a lineage that had been recently introduced into Brazil subsequently adapted to the Brazilian surgical niche (Figure 50), although it is not possible to definitively rule out the presence of the LCA of the epidemic lineage in the Brazilian environment. However, taking together the close relationship with the DCC, the recent emergence of the lineage and the knowledge that poor sterilization practices were observed in hospitals where outbreaks occurred, adds further weight to a previously suggested hypothesis that an *M. a. massiliense* lineage was introduced into the Brazilian hospital niche, exposed to non-lethal levels of GTA which led to the selection of a GTA tolerant lineage which subsequently spread throughout hospitals in Brazil (131, 373, 374, 376, 378).

How the epidemic lineage has spread across Brazil remains unclear. Contamination of aquatic environments, the movement of surgeons with their own equipment and the contamination of inactivated GTA have been proposed as possible mechanisms (373, 377). Drawing a definitive conclusion to this question is not possible due to the lack of any further epidemiological information for the samples used in this study. However, the genetic data does provide evidence suggesting the lineage was transmitted to the outbreak locations throughout Brazil as the phylogeny of the Brazilian lineage shows evidence of geographical and temporal structure (Figure 48B). This is more consistent with the idea of transmission than dissemination from a point source, which would be represented by a star shaped phylogeny with no geographical substructure. Furthermore, point source outbreaks would be expected to occur contemporaneously, thus, if a single source of contaminated GTA

distributed to the outbreak locations was responsible the outbreaks in different states would have been expected to occur at the same time.

Whilst there is strong evidence that the lineage has been transmitted to the differing outbreak locations, the vehicle enabling the transmission is unclear. Evidence provided by the MSTs (Figure 51), which suggest that transmission events from Pará seeded the outbreaks in states as far away Rio Grande do Sul, which is approximately 4000km south of Pará, and shows a high level of genetic similarity between the outbreak sub lineages responsible, imply that it's highly unlikely that the lineage spread from differing outbreak locations through aquatic systems as greater genetic diversity would be expected and it would require water systems linking very distant locations. Furthermore, it's highly likely outbreaks not associated with video-assisted surgeries would be expected if contamination of hospital water supplies was involved in the dissemination of the epidemic lineage (376). Therefore, the genetic data fits best with the scenario of human mediated transmission of the epidemic lineage and combined with the evidence of surgeons moving between hospitals where outbreaks occurred and equipment from different laparoscopic surgical teams being cleaned together suggests that the movement of surgeons with contaminated equipment is likely to be responsible for the spread of the epidemic lineage (131, 373).

It was not possible, due to incomplete sampling, to predict the transmission event that led to every outbreak. However, the MSTs did indicate that Pará acts as the main hub for transmission (Figure 51), as well as highlighting the role of this state as a continued source of the lineage throughout the duration of the epidemic and despite the outbreak in Pará ending in 2005 (374). This suggests that the lineage is capable of persisting in the hospital niche for significant periods of time, as was also observed in Rio Grande do Sul, and emphasizes the importance of maintaining sterilization practices and the risk of further outbreaks if these are not enforced (375). The combination of the ability to persist in the hospital environment and undergo long distance spread, as is evident by the transmission of the lineage between geographically distant states in Brazil, also has serious implications for the global spread of *M. a. massiliense* lineages.

Phenotypic and molecular analyses have shown that the Brazilian epidemic lineage had begun to adapt to the Brazilian surgical environment, with the lineage found to be more virulent than *M. a. massiliense* CIP108297 as well as GTA resistant (131, 379). Previous WGS analysis showed unique indels occurred in the genomes from the Papworth pulmonary infection outbreak and Brazilian post-surgical wound infection epidemic which could also

suggest the lineage has adapted to a specific niche (89).The smaller genome size of the Brazilian epidemic lineage is potentially further evidence of this adaptation (Figure 53). Specifically, it could be evidence of reductive evolution, which has been observed in other Mycobacterial pathogens, *Mycobacterium leprae* and *Mycobacterium ulcerans,* although on a much longer time scale (205, 390). This process has been observed on a comparable time scale in the pathogens *Salmonella enterica* subsp. *Enterica* serovar Typhimurium and *Salmonella enteritidis* (391, 392). On the other hand, the median genome size of the Brazilian epidemic lineage falls between the 5th and 95th percentile of the genome sizes for *M. a. massiliense*, and therefore it could be argued that the reduction in the genome size of the Brazilian lineage does not exceed the variation expected between *M. a. massiliense* lineages and consequently could suggest that this is not evidence of reductive evolution. The fact that only two of the 14 deletions observed in the Brazilian lineage were unique to this lineage (Figure 52) could be seen as further evidence that reductive evolution hasn't occurred.

The function of the deletions incurred by the Brazilian lineage were predicted using GO-term analysis. Del_12078_1#71_02620_02653 had 13 CDSs with functions related to oxidation-reduction processes and nine CDSs with functions associated with metabolic processes, suggesting that it may play a role in the degradation of an environmental metabolite, which in turn could be seen as evidence that the Brazilian lineage is losing genetic material it no longer requires in its new niche (Appendix table 4.3, table 4.4). However, experimental analysis is required to conclusively determine the function of both the deletions that occurred uniquely in the Brazilian lineage and in turn gauge the strength of the evidence they provide in terms of how the Brazilian lineage is adapting to its new niche.

The gain of genetic material is also a mechanism by which bacteria adapt to new niches. Two plasmids, pMAB01 and pMAB02, are harbored by the majority of isolates associated with the post-surgical wound infection epidemic in Brazil and only a few other isolates in the MABSC global population (Figure 48B, Figure 54A), suggesting these plasmids may be playing a role in the adaption of this lineage. There is evidence that pMAB01, which is conjugative and encodes resistance to antibiotics, mercury and quaternary ammonium compounds, was lost in specific geographical regions (Figure 48B) (87). However, this is probably due to loss of the plasmid during culture as the isolates from Goiás are known to have harbored pMAB01 in previous analysis and spontaneous curing of pMAB01 has been observed in Brazilian lineage isolates (130, 373, 377). pMAB02 was found to be encoded by all the Brazilian isolates sequenced in this study and just five other isolates in the global

population (Figure 54A). This mobile element has been identified from the Brazilian lineage before, both as Mycobacteriophage Adler and in all likelihood the three uncharacterized contigs in the assembly of CRM-0020 made up of mainly hypothetical proteins described by Tettelin et al. 2014 (91). However, firstly the gene content, such as the presence of plasmid mobility genes *traI* and *traM* and plasmid replication genes *parA* and *parB* (Figure 54B) and secondly, the read pair evidence, suggest the second contig in BRA_PA_42 is a circular plasmid. Furthermore, its co-occurrence with pMAB01, which encodes a complete type IV secretion system and has been shown to be conjugative, suggests it is potentially mobilizable (87, 393).

The function of both pMAB01 and pMAB02 in the adaption of the Brazilian lineage has not been established. pMAB01 encodes multiple resistance genes suggesting that it could play a significant role in an opportunistic pathogen, although it has been shown to not be responsible for the GTA resistant phenotype of the Brazilian epidemic lineage. It was only possible to predict a function for 32 of the 121 CDSs encoded by pMAB02, although it was found to encode a complete T7SS (Figure 54B).

T7SSs play a critical role in the pathogenicity of *M. tuberculosis* as well as other Mycobacterial pathogens (394). There are five types of chromosomally encoded T7SSs, labelled ESX-1 to ESX-5, encoded by various diverse Mycobacteria. ESX-1 and ESX-5 specifically are associated with host-pathogen interactions and escape from the phagosome, whilst ESX-3 is involved in iron and zinc acquisition (394-398). However, plasmid encoded T7SS have only recently been described (387, 388). Whilst there is nearly complete gene synteny between the T7SS encoded by pRAW, a plasmid harbored by *M. marinum* (Figure 54C), the AAI between the T7SS on pMAB02 and the others shown in Figure 54C ranged between 36% and 38% suggesting they diverged from one another a significantly long time ago. The T7SS encoded by pRAW has been shown to play a part in a novel conjugation mechanism, which could be a possible function of the T7SS encoded by pMAB02, although pMAB02 could also be mobilizable due to the presence of pMAB01, as previously discussed (387). It is also possible, given the increased virulence and GTA resistance phenotypes displayed by the Brazilian epidemic lineage that the T7SS system is playing a role in the expression of these phenotypes (379). On the other hand, the presence of pMAB02 in five other MABSC global population isolates (Figure 54A) suggests that, particularly in the case of the increased virulence phenotype, this is unlikely as the DCC closely related to the Brazilian epidemic lineage has also been shown to display increased virulence and only two

of the DCC isolates harbour pMAB02 (73). Experimental analysis is required to fully determine the role pMAB02 is playing in the adaptation of the Brazilian epidemic lineage.

The ability of the Brazilian epidemic lineage to survive exposure to GTA provided it with the selective advantage that enabled it to thrive in the Brazilian surgical niche. However, the molecular mechanism responsible for its resistance is unknown. Whilst pMAB01 has been shown not to be responsible for this phenotype, despite the presence of a known biocide resistance gene, no other analysis has been done to find the cause of this phenotype (87). With the current understanding of the gene content of pMAB02 it is difficult to argue that it is responsible for the GTA resistance phenotype. Therefore, changes in the chromosome were investigated as possible candidates to explain the GTA resistance and specifically porins because they have been associated with GTA resistance in *M. chelonae* (131, 287). The initial observation of the deletion of a porin gene in 180 of 189 Brazilian epidemic lineage isolates (Figure 55) showed a change in the porin region had occurred and suggested, given that porins were responsible for GTA resistance in *M. chelonae*, they potentially could be responsible in this lineage of *M. a. massiliense*. On the other hand, given the GTA resistant phenotype was evident in all the Brazilian epidemic lineage isolates tested, the fact that nine isolates from this lineage were found to encode both *mspA* and *mspB* (Figure 55) hinted that potentially the porins were not responsible. This was shown to be the case when the growth of *M. a. massiliense* CIP108297 WT, CIP108297 *mspA* KO, CIP108297 *mspB* KO and the Brazilian lineage isolate, CRM-0019, in 0.2% and 2.2% GTA solution was tested and all the isolates bar the epidemic isolate from Brazil were found to be susceptible (Figure 57). Therefore, porins are not responsible for the GTA resistance phenotype in the Brazilian epidemic lineage and thus either pMAB02 or a different chromosomal change must be responsible for this phenotype.

5.5 Conclusions and Future Directions

In summary, WGS of isolates associated with the post-surgical wound infection epidemic in Brazil has shown that a single lineage of *M. a. massiliense* is responsible. The lineage is closely related to a DCC, which when taken together with the estimated emergence of the epidemic lineage in 2003 and the poor sterilization practices observed suggests that a lineage recently introduced into Brazil was exposed to non-lethal levels of GTA, adapted to become GTA resistant and subsequently spread throughout hospitals in Brazil, with the genetic data suggesting that the lineage was transmitted to these locations as opposed to being spread through water systems or contaminated inactivated GTA. Large chromosomal deletions and the presence of a second novel plasmid suggested that the lineage has

adapted to a novel niche, although further experimental analysis is required to link these observations to the known virulence and GTA resistance phenotypes of the epidemic lineage, with a deletion of a porin shown not to be responsible for the GTA resistance phenotype.

The evidence presented here of the rapid adaptation of globally circulating MABSC lineage to a specific novel hospital niche and its subsequent transmission to multiple geographically distant locations highlights the threat posed by the MABSC

6. Conclusions

6.1 A restatement of research aims

The *Mycobacterium abscessus* species complex (MABSC) has emerged as an increasingly common opportunistic pathogen, with particularly serious consequences for people with underlying lung conditions such as Cystic Fibrosis (CF). The highly antibiotic resistant nature of the MABSC and the toxicity of the currently available treatment means that the outcome of MABSC infections are poor. Thus novel antibiotics are urgently needed.

MABSC whole genome sequencing (WGS) analysis has already significantly enhanced our understanding of these opportunistic pathogens by showing that the majority of infections in people with CF are caused by isolates that form part of a few recently emerged, more virulent and globally disseminated lineages (Figure 6), as well as uncovering evidence that the MABSC is capable of indirect person-to-person transmission (70, 73).

This thesis aimed to build upon this, with the broad aim of using WGS to investigate what genetic changes have occurred as the MABSC has evolved during its emergence as a more common opportunistic pathogen. Through such analyses our understanding of the adaptive mechanisms and functional pathways the MABSC is using to survive and thrive in the human host could be improved and potentially novel drug targets could be uncovered.

6.2 Findings with clinical and epidemiological implications

6.2.1 The emergence of the most prevalent MABSC lineages is due to increased opportunity as opposed to the acquisition of single common genetic factor.

Particular lineages of the MABSC have been shown to be responsible for the majority of MABSC infections in people with CF (73). However, whether the same genetic factors were responsible for the emergence of these lineages had not been investigated. Through dN/dS, SNP density and pangenome analysis, no evidence was uncovered to suggest the same genetic factor had driven the emergence of the three most prevalent lineages. This suggested that a host factor rather than a single bacterial factor could be the common driver of their emergence and the increasing availability of the CF niche due to the increasing number of people with CF living longer is the most likely candidate. Each of the dominant circulating clones (DCCs) were found to encode a unique combination of potential virulence factors that could explain the increased virulence of these lineages and why it was these lineages that had expanded to the greatest extent as opposed to other MABSC lineages. However, the fact that the rapid emergence and global spread of these more virulent

lineages was likely to be due to opportunity as opposed to a common genetic factor suggests that there should be an awareness (and maybe even an expectation) of the potential for uncommon opportunistic pathogens to rapidly emerge and spread amongst the CF community, even when there is limited evidence to suggest the particular species is capable of doing so.

The analysis in this thesis did not extend to comparing the dN/dS, SNP density and gene content of all clustered lineages in comparison to all unclustered lineages. This analysis should be performed in the future and could uncover further functional gene content that could have contributed to the emergence of the more prevalent MABSC lineages. However, the evidence thus far suggests that it is unlikely that a single combination of genetic factors that defines an epidemic lineage of the MABSC exists.

6.2.2 Detection of potential novel MABSC antibiotic resistance variants

Variants had accumulated over time within the host in four genes which are known to be linked to antibiotic resistance: 16s rRNA, 23s rRNA, *erm(41)* and *eis2* (77, 152, 313, 343). Whilst many of the variants accumulated by 16s rRNA, which cause aminoglycoside resistance, and 23s rRNA, which cause macrolide resistance, had been previously observed, novel antibiotic resistance variants were uncovered. These require experimental validation to determine whether they are responsible for antibiotic resistance.

Much less easy to understand is what the likely impact of the variants accumulated by Eis2 is. Eis2 has recently been shown to be linked with the high level of intrinsic amikacin resistance displayed by *M. abscessus* and to be part of the *whiB7* regulon (152). Thus it is unclear how the accumulation of nonsynonymous variants would be beneficial if they cause loss of function in this gene, deletion of which has been shown to be increase the susceptibility of *M. abscessus* to aminoglycosides (399). Therefore a change or adjustment in function seems a more logical consequence of the accumulation of these nonsynonymous SNPs. On the other hand, given the propensity for pathogens adapting to chronic infection to adapt to become less virulent, if *eis2* is also performing a virulence associated function, it could be that the loss of its function is beneficial to long term survival in the host and that its intrinsic aminoglycoside resistance function could be compensated for by an acquired aminoglycoside resistant mutation in 16s rRNA (369). Despite this uncertainty, the data provided here should provide a useful resource to further investigate the role of this gene.

This analysis also highlights how WGS can be applied to detect mutations potentially associated with antibiotic resistance.

6.2.3 WGS of the Brazilian post-surgical wound infection epidemic shows that the MABSC is capable of causing large scale outbreaks, persisting in the hospital environment and being transmitted long distances.

The largest number of non-CF related MABSC infections has been observed in Brazil where an epidemic of post-surgical wound infections caused by a glutaraldehyde tolerant lineage of *M. a. massiliense* has been ongoing since 2004 (130). There had been uncertainty over whether a single lineage was responsible and how the lineage had spread to geographically distant locations throughout the country (89). WGS of 188 isolates from the epidemic showed that a single lineage of *M. a. massiliense* that had recently been introduced into Brazil adapted to become glutaraldehyde (GTA) tolerant and then spread through several waves of transmission to geographically distant regions of Brazil.

The scale of this epidemic, with over 2000 cases recorded, highlights the ability of the MABSC to cause large scale nosocomial outbreaks. This is concerning given the ability displayed by this lineage to rapidly acquire adaptations beneficial to its novel environment, to persist in the environment, as displayed by the fact that transmission events continued to emanate from Belém even after the outbreak ceased, and to survive long distance transmission events. These features suggest that there is the potential for this lineage to be introduced and cause outbreaks in health care settings around the world. Although, it should be noted that outbreaks in Brazilian hospitals ceased when new surgical tool disinfection protocols were introduced and thus it is unlikely that outbreaks will occur unless these are inadequate (as was witnessed by further outbreaks in Brazilian hospitals when implementation of the new disinfection protocols lapsed). There has yet to be evidence that this lineage has spread outside of Brazil and caused further outbreaks. Although, a point source outbreak caused by the dissemination of contaminated ultrasound gel in Taiwan was found to be closely related to the Brazilian epidemic lineage through multi locus sequence typing (400). It would be interesting to WGS representatives from this outbreak to establish their context in the MABSC global population and examine how closely related the causal agents of these outbreaks are.

Evidence from both the global dissemination of the DCCs and the spread of the Brazilian epidemic lineage to geographically distant parts of Brazil show that some lineages of the MABSC are capable of surviving long distance transmission events. Whilst there is evidence

that the mechanism for local transmission of MABSC infections can be *via* fomites or long-lived cough aerosols, how the Brazilian lineage was transmitted long distances within Brazil and how the DCCs have been globally disseminated remains unclear (73). Given their differing epidemiology, it is likely their vehicles of transmissions are different. The lack of epidemiological information meant that it was not possible to clarify what vehicle had enabled the spread of the Brazilian epidemic lineage. A previously stated hypothesis, that the WGS data supports, is that the movement of surgeons with contaminated equipment is responsible (131). How the DCCs have become globally disseminated is less clear. Potential areas where further research may contribute to the understanding of the spread of the DCCs would be a greater understanding of the role of asymptomatic carriage in the epidemiology of the MABSC as well as a greater understanding of the distribution of the MABSC in the environment.

This analysis also highlights how WGS can be useful in identifying whether isolates from a suspected outbreak are monophyletic and can also distinguish whether an outbreak is from a point source or not, but it also emphasizes that without sufficient epidemiological information the full power of WGS as an epidemiological tool cannot be realized.

6.2.4 The loss of an Msp family porin does not explain the glutaraldehyde resistant phenotype associated with the Brazilian epidemic lineage

The Brazilian epidemic lineage rapidly adapted to the surgical niche in Brazil by becoming GTA resistant (131). Mycobacterium smegmatis porins (msp) had previously been associated with GTA resistance in *M. chelonae* and *M. smegatis* and the observation that an msp porin at the same locus had been deleted in the majority of isolates from the Brazilian epidemic lineage led to the hypothesis that porins could be responsible for the GTA resistance displayed by this lineage (287). However, this was found not to be the case, similarly to *P. fluorescens* (401), and thus the cause of the GTA resistant phenotype remains unclear. As GTA remains one of the most widely used disinfectants around the world, due to its low cost and lack of volatility, and resistance has been observed in multiple nosocomial pathogens, including the CF pathogen *P. aeruginosa*, understanding how organisms become resistant to GTA is essential to optimize its usage (402). Possible avenues for further research include the role of efflux pumps and, given the strong evidence that GTA's mechanism of action involves forming cross links with cell wall associated proteins, genetic changes affecting these seem worthy of further investigation (287, 403). As the deletion of a porin was not found to be responsible for the GTA resistant phenotype it is possible that it is

playing a role in other phenotypic characteristics of the Brazilian epidemic lineage, such as its increased virulence (379).

6.3 Findings which contribute to our understanding of how MABSC is adapting to cause disease

6.3.1 Changes in common pathoadaptive pathways could have contributed to the emergence of the most prevalent MABSC lineages

All three DCCs were shown to have undergone changes, either whilst the LCA of each lineage existed in its natural habitat or that were selected for in the early selective sweeps prior to the clonal expansion of each lineage, that could have contributed to the emergence of these lineages as the most common cause of MABSC infections in CF. There were examples amongst the candidates of genes which participate in pathways that have been linked to the adaptation of other organisms to survival within the host, which suggests that the MABSC is adapting *via* similar mechanisms to other pathogens. Interestingly, recent publications also investigating the functional adaptation of the MABSC uncovered similar evidence of overlap and came to similar conclusions (181, 404).

Examples from this analysis include i) the presence of genes involved in the transport, catabolism and biosynthesis of branched chain amino acids (BCAA) in DCC1, which have been shown to be important to the virulence of *M. tuberculosis, S. pneumoniae* and *Staphylococcus aureus* (244-246), ii) the presence of the same cluster of genes associated with ubiquinone biosynthesis by DCC1 and DCC2 which has been linked with virulence in *E. coli* (258, 282) and perhaps most significantly iii) the observation of changes in β-oxidation of fatty acids in all three DCCs. Significantly, genes associated with the β-oxidation of fatty acids genes were also shown to be upregulated by *M. a. abscessus* ATCC19977 when grown in Synthetic Cystic Fibrosis Medium (SCFM) (181). β oxidation of fatty acids has been shown to be important to the pathogenesis of both *M. tuberculosis,* which switches to β oxidation of fatty acids as a carbon source during survival within the phagosome, and *P. aeruginosa* which has been shown to upregulate these genes when grown in SCFM (284, 405).

There were limitations to this analysis, which include the possibility that the functions encoded by these genes are present within the other less prevalent MABSC lineages or that the candidates could have emerged through a limitation of pangenome analysis that makes

correctly splitting paralogous genes challenging. However, it is not possible without experimental analysis to determine whether other genes are providing these potential beneficial functions in other MABSC lineages. Their overlap with known virulence pathways in other pathogens supports these as strong candidates for further investigation in the bid to understand how the MABSC is adapting to become a more prevalent opportunistic pathogens.

6.3.2 DCC3 acquired a methyltransferase which could be contributing to its increased ability to survive intracellularly.

A methyltransferase encoded on a mobile element was found to be present in all DCC3 isolates. Through PacBio analysis the methyltransferase was shown to modify an RGATCC motif. Phagocytic uptake and intracellular survival assays suggested that the methyltransferase was potentially playing a role in the intracellular survival of the DCC3 lineage. Whilst these results are preliminary, as the phenotype has not been able to be reproduced, this could potentially be the first evidence that a change in the methylome of an MABSC lineage is associated with its virulence. Indeed, changes in the methylome have been linked with virulence in other organisms (305, 406). If this phenotype is confirmed, this would suggest a wider analysis of the methylome of MABSC lineages with differing degrees of virulence and transmissibility could potentially be worthwhile. This was also the first attempt to functionally validate a candidate associated with the emergence of the most prevalent MABSC lineages and thus, even though this result is preliminary, it suggests that the methods used to address the aims set out in this thesis have potentially successfully identified valid candidates.

6.3.3 Mce operons are implicated in both the emergence and ongoing expansion of the DCC:

DCC3 was found to have acquired a complete *mce* operon prior to, or in the early selective sweeps prior to, its clonal expansion. The regulator of an *mce* operon was found to have accumulated a significant number of nonsynonymous SNPs during the ongoing clonal expansion of DCC1. These results suggest that these known virulence factors are contributing to the adaptation of the MABSC (176, 259, 407). Thus these seem promising candidates for further investigation. Analysis showed that both MABSC clustered and unclustered isolates on average encoded the same number of *mce* operons but pangenome analysis suggested that these were not all part of the core. Therefore, this suggests that the differing lineages may encode paralogous *mce* operons with differing functions which could

be contributing to their differing degrees of prevalence in disease. If the two *mce* operons were found to have differing functions a further layer of intrigue would be added, given that one candidate *mce* operon was identified as associated with predisposing DCC3 to be successful in the CF lung environment and the other was associated with the ongoing spread of DCC1, and thus their functions could be associated with the different stages of adaptation of an MABSC epidemic lineage.

6.3.4 Mycobacterium abscessus is adapting to the CF lung through routes commonly used by CF pathogens

Similarly to the adaptive mechanisms that were linked with the emergence of the DCCs, the MABSC was shown to use many of the same mechanisms as other as other CF pathogens to adapt to the CF lung, such as the acquisition of a hypermutator phenotype, changes in cell wall associated genes, global transcriptional regulators and antibiotic resistance associated genes (39, 40, 308). Experimental analysis is needed to understand the role of these genes but the virulence associated functions of many of their orthologs suggest that they could be playing a significant role in the pathogenesis of the MABSC. Whilst intriguingly some of their orthologs have also already been proposed as novel drug targets which suggests they could also be suitable targets for novel drugs for the MABSC (371, 372).

Whilst these candidates could have highlighted key virulence factors in the MABSC, through protein sequence analysis, some of the variants accumulated within these genes were shown to have incurred nonsynonymous changes both at sites known to participate in the formation of the correct topology of the protein and in regions binding DNA or their co-factors. Thus this dataset could be of use to those investigating the structure of these proteins.

6.3.5 PhoR is a key virulence factor in the MABSC

PhoR stood out from the other candidates that accumulated nonsynonymous SNPs over time within the host due to both the number of changes it accumulated and the fact they were clustered in the sensor loop. Imperfect evidence from RNA-seq analysis suggested the PhoPR regulon was controlling the expression of some genes with similar functions to that of the PhoPR regulon in *M. tuberculosis*. The functional impact on the protein of these SNPs remains unclear, with two hypotheses suggested: i) the accumulations of variants increase the efficiency of PhoR to sense its environmental cue and thus *M. abscessus* is better able to rapidly adapt to changing environmental conditions within the host or ii) as CF pathogens tend to become less virulent as they adapt to cause chronic infection, the accumulation of

variants could result in a decrease in the function of PhoR, and thus a reduction in virulence. In *M. tuberculosis,* the accumulation of variants in the PhoR sensor loop results in the attenuation of virulence, which supports the second hypothesis. Either way, the evidence presented in this thesis suggests that PhoR is a key virulence factor in the MABSC.

6.3.6 A second plasmid, pMAB02, which encodes a type VII secretion system, is present within the Brazilian epidemic lineage

Type VII secretion systems, which play a role in the virulence of *M. tuberculosis*, could also be playing a role in the virulence of MABSC (396). pMAB02 the a novel plasmid identified through this thesis as present in the Brazilian epidemic lineage. The Brazilian epidemic lineage has been shown to have increased virulence and a genetic explanation for this has not been proposed (379). Thus the type VII secretion system encoded by the plasmid may be of interest. The plasmids associated with the MABSC are also relatively understudied, and if further research into pMAB02 or indeed pMAB01 show that plasmid encoded genes are contributing to the success of this epidemic lineage, the role of plasmids in the MABSC could become of significant interest.

6.4 Future Directions

Much of the research in this thesis was exploratory, with the purpose of generating hypotheses that could explain how the MABSC has evolved as it has become a more prevalent opportunistic pathogen. Whilst follow up analysis of several candidates has begun, a significant amount of further research is required to validate the candidates identified. These results will contribute to gaining a clearer picture of how the MABSC is able to cause disease and could potentially indicate potential novel drug targets.

Ongoing further analysis has already shown that the porins undergo several deletions over time within multiple patients and isolates encoding a single porin were subsequently shown to be more virulent than those encoding both. This suggests that the loss of the porin in the Brazilian epidemic lineage could be associated with its increased virulence. Further research is also already ongoing regarding the role of PhoR, with evidence thus far supporting the hypothesis of the accumulation of variants in the sensor loop is associated with a reduction in virulence and therefore adaptation to persistence. This emphasizes even more the need to clarify the *M. abscessus* PhoR regulon as it suggests the genes under its control are associated with virulence.

As this dataset is publicly available it provides the opportunity for the MABSC to continue to be investigated through genomic analyses. There are many remaining challenges and questions. This thesis addressed its aims in the knowledge that some types of genetic variation that could be contributing to the evolution and adaptation of the MABSC have not been investigated. Perhaps most significantly the influence of recombination is not thoroughly addressed. Both Tan et al. (2017) and Sapriel et al. (2016) emphasize the role of recombination in the evolution of the MABSC, and find that highly admixed MABSC are more commonly associated with causing infections in CF (86, 90). This conclusion would benefit from being placed in the global population context afforded by this data set. Analysis from this thesis, showing that many of the genes that defined the DCC lineages occurred in blocks, suggests that large scale recombination events are contributing and thus further research into MABSC recombination is warranted.

A significant gap remains in our knowledge with regards to how the MABSC is able to be transmitted. A greater understanding of the environmental ecology of the MABSC would enhance our understanding of the global distribution of MABSC lineages could in turn contribute to our understanding of how they have spread globally. Avenues of research such as asymptomatic carriage may also be worthy of further investigation. Furthermore, a greater availability of environmental MABSC isolates would enhance our ability to detect the key functional pathways the MABSC uses to cause disease.

6.5 Closing comments

Through this thesis, the functional genomic changes that could have contributed to the emergence, ongoing spread and host adaptation of MABSC lineages have been explored. The findings have highlighted functional pathways used by the MABSC to cause disease in the human host and have uncovered promising candidates for further research.

7. Materials and Methods

This chapter summaries the dataset and methods used in multiple chapters in this thesis. The collection of the metadata used in this dataset was orchestrated by Andres Floto and Julian Parkhill. Clinical metadata was collected by Dorothy Grogono. Josephine Bryant performed the quality control analysis on the WGS dataset.

7.1 Datasets

7.1.1 Global population dataset

The main dataset used in this analysis was a combination of the datasets published by Bryant et al. in their analyses of the MABSC global population structure (73) and of the outbreak of *M. abscessus* at Papworth hospital in the UK (70). This dataset consisted of Clinical MABSC isolates submitted from all the major Cystic Fibrosis (CF) centers in the UK and the five mainland mycobacterial reference laboratories: National Mycobacterial Reference Laboratory (NMRL); Regional Centre for Mycobacteriology, Birmingham; Regional Centre for Mycobacteriology, Newcastle; Scottish Mycobacteria Reference Laboratory and Wales Centre for Mycobacteriology. Isolates were also contributed from European CF centers in the Republic of Ireland (Cystic Fibrosis centers from St Vincent's Hospital Dublin), Sweden (Gothenburg), Denmark (Copenhagen and Skejby) and the Netherlands (Nijmegen) as well as from the USA (University of North Carolina Chapel Hill) and Australia (Queensland).

The global population dataset was supplemented by 29 publicly available isolates, including isolates from France, China, Malaysia, South Korea and Brazil and further isolates from the USA, including representatives from the Seattle CF center outbreak (119), and the UK (appendix table 5.2). These provided an additional eight *M. a. abscessus* isolates, two additional *M. a. bolletii* isolates and 19 additional *M. a. massiliense* isolates.

The final dataset, once contaminants had been removed, consisted of 1252 MABSC isolates, with 781 *M. a. abscessus* isolates, 108 *M. a. bolletii* isolates and 363 *M. a. massiliense* isolates. These isolates were obtained from 525 patients (appendix table 5.1). Including publicly available isolates, the final dataset consisted of 1281 MABSC isolates, with 789 *M. a. abscessus* isolates, 110 *M. a. bolletii* isolates and 382 *M. a. massiliense* isolates. These isolates were obtained from 553 patients (appendix tables 5.1 and 5.2).

7.1.1.1 Single isolate per patient dataset

In chapters 2, 3, 5 phylogenetic trees were constructed using a single representative isolate for each patient. This dataset included 525 isolates marked in appendix table 5.1, the *M. a. abscessus* ATCC19977 reference genome and the 29 publicly available isolates (appendix table 5.2).

7.1.1.2 Within host evolution dataset

In Chapter 4 the genetic changes occurring over time within a patient were investigated. The MABSC global population dataset included multiple isolates for 201 patients (those used in the analysis are marked in appendix table 5.1, for more information see chapter 4).

7.1.2 Brazilian post-surgical wound infection epidemic dataset

A total of 190 isolates were obtained from nine states in which outbreaks of SSTIs, caused by the MABSC, occurred in Brazil. The analysis of this dataset is reported in chapter 5.

7.2 DNA extraction and whole genome sequencing

All culturing and DNA extraction steps for the sequencing of isolates analysed in this thesis were carried out by external collaborators. The methods used to extract the DNA from the global population dataset isolates are summarized in (73). The methods used to culture and extract the DNA from the Brazilian dataset are described in chapter 5. Library preparation and sequencing of all the isolates was carried out by the DNA pipeline team at the Wellcome Trust Sanger Institute. All isolates were subjected to multiplexed paired end sequencing on either the Illumina HiSeq (either 2000 or 2500 technology) or MiSeq platforms, resulting in the production of raw read files in fastq format.

7.3 Mapping and variant calling

Mapping and variant calling were carried out using an in house pipeline (developed by Simon Harris). Raw reads were mapped, where appropriate, to the MABSC subspecies reference genomes: *M. a. abscessus* ATCC19977, *M. a. massiliense* CIP108297 and *M. a. bolletii* BD (table 15) (81, 82, 86). Mapping was carried out using BWA-MEM (v. 0.7.12-r1039), with default parameters, resulting in the generation of BAM files (215). PCR duplicate reads were identified and marked in the BAM files using Picard MarkDuplicates (v.1.127). GATK indelrealigner (v.3.4-46-gbc02625) was used to realign reads locally around indels.

Table 15: Summary of the reference genomes used in this thesis[13]

	M. a. abscessus	***M. a. massiliense***	***M. a. bolletii***
Reference name	ATCC19977	CIP108297	BD
Type	Finished	Draft	Draft
Total chromosome length (bps)	5,067,172	4,890,609	5,048,007
Total number of contigs (chromosome)	1	80	22
Total number of CDS	4,920	4,821	4,903
Total number of RNAs	50	71	77
NCBI accession number	NC_010394.1	GCA_001792625.1	AHAS00000000.1

Samtools mpileup (v. 1.2.1) and bcftools call (v. 1.2.1) were used to determine the consensus base call from the aligned reads (216). Default parameters were used for Samtools mpileup apart from limiting maximum per sample read depth (-d) and the per sample read depth for indel calling (-L) to 1000, setting the adjusted mapping quality score (-C) to 50 in order to downgrade the quality scores of bases called from reads with a large number of mismatches, and selecting a threshold of 8 for the minimum number of gapped reads required for an indel to be considered (-m). Bcftools was then used to call the bases. All alternative alleles were recorded (-A), masked reference sites (-M, Ns) were kept and a prior (mutation rate, -P) of 0.001 was applied. Finally, filters, which were estimated to reduce the number of false positive base calls to less than 1 SNP per genome, were applied (208). The following criteria had to be met in order for a base to be called: a SNP/Mapping quality ratio of 0.8; a minimum depth of 8 reads supporting the called base, including at least 3 reads supporting the called base on each strand; a minimum base quality score of 50; a minimum mapping quality score of 20 and p-values less than 0.01 for strand bias, mapping bias and tail distance bias. This resulted in the creation of a pseudo sequence for each isolate, with base calls which did not meet the above criteria recorded as an N.

[13] The contigs of the draft genomes for *M. a. massiliense* CIP108297 and *M. a. bolletii* BD were reordered against the *M. a. abscessus* ATCC19977 reference genome and re-annotated using Prokka.

7.4 Extracting variant positions from alignments and constructing phylogenies

Variant positions were extracted from alignments of multiple isolates using SNP-sites (v.2.3.2) (217). For phylogenetic analyses only sites for which greater than 50% of isolates had a base called (i.e not Ns) were included. Maximum likelihood phylogenetic trees were inferred from these alignments via RAxML (v.8.2.8), using the Generalised Time Reversible (GTR) model of evolution and the GAMMA model for among site rate variation (61). 100 bootstrap replicates were performed.

On the majority of occasions it was useful to map the variants back onto the phylogeny, which both allowed for the branches to be scaled by number of SNPs and for ancestral changes to be examined within the context of the phylogeny. This was achieved using an inhouse script (developed by Simon Harris) which reconstructs the ancestral sequence for each node of the phylogeny using an ACCTRAN (ACCelerates the evolutionary TRANsformation of a character) parsimony model (218).

7.5 Phylogenetic clustering

The clustering method, TreeGubbins, used by Bryant et al. (2016) in their analysis of the MABSC global population structure, was also applied in this analysis (73). TreeGubbins (developed by Simon Harris) is an algorithm that allows you to determine significantly dense nodes in a phylogeny enabling the differentiation of densely clustered isolates from more loosely clustered isolates, which could indicate recently emerged lineages. TreeGubbins, using a one-dimensional scanning statistic likelihood function, calculates the density of each node (the mean descendant branch length) and compares it to the expected density of each node (the mean branch lengths of the remaining tree). The statistical significance of each node, starting with the node with the maximum likelihood, is determined by randomly re-assigning, 100 times, the branch lengths across the tree and recalculating the density of each node. The process is repeated until no significant nodes (P-value < 0.05) are detected.

This clustering method was applied to the subspecies phylogenies, consisting of a single isolate per patient and publicly available isolates, inferred from i) an alignment of the variant positions after all the isolates were mapped to the *M. a. abscessus* ATCC19977 reference genome and ii) from the alignments of the variant positions after the isolates were mapped to their respective subspecies reference genomes, either *M. a. abscessus* ATCC19977, *M. a.*

bolletii BD or *M. a. massiliense* CIP108297. Only clusters with greater than five isolates (to provide enough power) and including isolates from multiple sites (to rule out clusters of isolates acquired from the same environmental source) were included in downstream analysis.

7.6 *De novo* assembly

De novo assemblies for all the isolates were constructed using an in house pipeline (408). This pipeline generates multiple assemblies, using Velvet (v.2.2.5) and VelvetOptomiser (v.1.2), for each isolate using differing kmer lengths (kmer lengths are varied between 66%-90% of the read length) (219). The assembly with the highest N50 (length of the longest contig, where half of the nucleotides in the assembly are in contigs of at least the length of that contig) is taken forward and improved using SSPACE to scaffold the contigs and GapFiller to fill in sequence gaps (409, 410).

7.7 Annotation

The *de novo* assemblies of all the isolates were annotated using Prokka (v.1.11) (220). Prokka predicts features encoded by the assemblies using Prodigal (CDSs) (411), RNAmmer (rRNA) (412), Aragorn (Transfer RNA genes) (413), SignalP (Signal leader peptides) (414) and Infernal (Non-coding RNA) (415). These features are then annotated in a hierarchical manner starting with searches against smaller more trustworthy databases before resorting to more general curated protein family databases, with an e-value of 10^{-6} as the cutoff for a significant hit.

7.8 Pangenome analysis

At times it was useful to compare gene content differences between groups of isolates. This was achieved using the pangenome software Roary (223). Only CDS of at least 120 nucleotides in length and with less than 5% of nucleotides unknown are included in the analysis. In the initial step, Roary clusters CDS iteratively with CD-hit starting with CDS which share 100% sequence identity and 100% match length. This is subsequently repeated with the thresholds being reduced by 0.5% each time until it reaches the cut off of 98%, which can be adjusted by the user. The core genes, defined as CDSs present in at least 99% of the isolates, are then removed and a representative sequence for the remaining CDSs are placed in a fasta file.

An all against all blastp is then performed between all the remaining CDSs, with a user defined threshold. The results of the all against all blast are then clustered using a Markov Cluster Algorithm (MCL), which produces the final result of clusters of homologous genes which are subsequently combined with the core genes identified via CD-hit. However, because these clusters can contain paralogous genes a further step is performed to try and reduce this using a conserved gene neighborhood method to detect orthologs amongst a group containing potential paralogs.

7.9 Functional enrichment analysis

GO-term enrichment analysis was used to determine whether candidate genes lists generated were enriched with particular functions. GO-terms had previously assigned to the CDSs encoded by the *M. a. abscessus* ATCC19977 reference genome (231). For candidates identified from the *de novo* assemblies, InterProScan was used to assign GO-terms to the Prokka annotated CDSs (416).

GO-term enrichment analysis was performed using the R package TopGO (v.2.20) (232). TopGO enables the user to apply one of several algorithms that take the GO hierarchy into account when testing whether a candidate list is enriched with particular MF, BP or CC using a Fisher's Exact test. Taking the GO hierarchy into account reduces the number of false positive as it reduces the impact of the inheritance problem. The inheritance problem refers to the problem of general GO-terms inheriting their more specific descendant GO-terms resulting in false positives (417). The weight01 algorithm, the default method in TopGO, was used in this analysis. Only GO-terms assigned to at least five genes were tested, to increase the power to detect enrichment. The p-values were corrected for multiple testing using the Benjamini-Hochberg method, with an adjusted p-value of 0.01 seen as significant.

Pathway analysis was carried using the Blast2GO (v.4.1.9) interface (233). The default settings selected by Blast2GO were used. Briefly, Blast2GO predicts the pathways a set of candidate genes are participating in by firstly comparing the candidate genes against the NCBI non-redundant sequence database. This is followed by the genes being assigned GO-terms. Each GO-term, where the information is available, is linked to an enzyme code. The enzyme codes are then compared against the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database to identify which pathways the candidate genes are participating in.

7.10 Detection of orthologous genes between MABSC isolates and other mycobacterial species

Often it was useful to determine whether a candidate gene was orthologous to a gene within another species of Mycobacteria, particularly the causative agent of tuberculosis, *Mycobacterium tuberculosis*, for which the most functional information is known. Orthologs between *Mycobacterium tuberculosis* H37Rv and the *M. a. abscessus* ATCC1977 reference genome have been previously identified and where possible were used (230).

In some cases the candidate genes identified were from the *de novo* assemblies and in such cases a reciprocal blast approach was used to identify orthologous genes. Local nucleotide blast databases were created from the genes, predicted by Prokka, to be encoded by either the *de novo* assembly of the MABSC genome of interest or the pangenome and for 4,032 genes encoded by the *M. tuberculosis* H37Rv reference genome (83). Each gene encoded by one genome was then compared using tblastx (v.2.2.25), with an e-value threshold of 0.00001, against the blast database of all the genes encoded by the other (386). For genes to be predicted as putative orthologs, the genes had to be each other's top hit, have a sequence identity greater than 50% and a match length of greater than 50%.

7.11 Detecting genes under selection

To identify genes under selection SNPs potentially acquired via recombination were required to be removed. The methods used to detect recombinant SNPs are described in the chapters were this method is applied.

Genes which had accumulated a greater number of nonsynonymous SNPs than would have been expected by chance were determined using a method based upon Ding et al's (2008) 'burden of mutation' approach (295). This method estimates the ρ_{SN} (synonymous mutation rate) by dividing the observed number of synonymous SNPs by the number of coding sequence bases in the reference genome (*M. a. abscessus* ATCC19977: 4,686,864bps; *M. a. massiliense* CIP108297: 4,514,468bps; *M. a. bolletii* BD: 4,648,845bps). The expected nonsynonymous mutation rate (ρ_{NS}) is then estimated using the following equation: $\rho_{NS} = \rho_{SN} \times R$. *R* represents the ratio of nonsynonymous sites to synonymous sites and is determined by permuting every base of every codon *in silico* and identifying whether it resulted in a synonymous or nonsynonymous change. This was done on a per gene level, with the number of synonymous SNPs accumulated per gene used to estimate the value of

ρ_{SN} per site per gene. If no synonymous SNPs were observed in a gene the synonymous mutation rate estimated for the whole genome was used. Finally, to obtain the expected number of nonsynonymous SNPs per gene, ρ_{NS} was multiplied by the gene length.

To determine if the observed number of nonsynonymous SNPs was significantly greater than the expected, a one tailed binomial test was used. The p-values were corrected for multiple testing using Benjamini Hochberg method. A p-value of less than 0.01 was seen as significant.

8. Appendix

All the tables summarized in these appendices are available on the accompanying CD.

8.1. Appendix for Chapter 2

Appendix table 1.1: Breakdown of the number the SNPs on the branches leading to the LCA of each of the DCCs and those occurring on the branches evolving completely independently of all the DCCs.

Appendix table 1.2: Summary of the 23 candidates genes identified through SNP density analysis as associated with the emergence of DCC1.

Appendix table 1.3: Summary of the 6 candidate genes identified through SNP density analysis as associated with the emergence of DCC2.

Appendix table 1.4: Summary of the 61 candidate genes identified through SNP density analysis as associated with the emergence of DCC3.

Appendix table 1.5: Summary of the functions of the genes flanking the regulator MAB_3565 (marked with a *) which was identified through SNP density analysis as associated with the emergence of DCC2.

Appendix table 1.6: Summary of the functions of the genes flanking the regulator MAB_3582 (marked with a *) which was identified through SNP density analysis as associated with the emergence of DCC2. ** represent a potentially frameshifted gene

Appendix table 1.7: Summary of the functions of the genes flanking the regulator MAB_4754 (marked with a *) which was identified through SNP density analysis as associated with the emergence of DCC2.

Appendix table 1.8: List of the GO-terms assigned to the 4,920 CDSs encoded by *M. a. abscessus* ATCC19977

Appendix table 1.9: Summary of the results of the Fishers exact test, performed with the R package TopGO, which examined whether the candidates identified through SNP density analysis as associated with the emergence of DCC1 were enriched with particular Molecular Function GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.10: Summary of the results of the Fishers exact test, performed with the R package TopGO, which examined whether the candidates identified through SNP density analysis as associated with the emergence of DCC1 were enriched with particular Biological function GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.11: Summary of the results of the Fishers exact test, performed with the R package TopGO, which examined whether the candidates identified through SNP density analysis as associated with the emergence of DCC1 were enriched with particular Cellular Component GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.12: Summary of the results of the Fishers exact test, performed with the R package TopGO, which examined whether the candidates identified through SNP density analysis as associated with the emergence of DCC2 were enriched with particular Molecular Function GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.13: Summary of the results of the Fishers exact test, performed with the R package TopGO, which examined whether the candidates identified through SNP density analysis as associated with the emergence of DCC2 were enriched with particular Biological function GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.14: Summary of the results of the Fishers exact test, performed with the R package TopGO, which examined whether the candidates identified through SNP density analysis as associated with the emergence of DCC2 were enriched with particular Cellular Component GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.15: Summary of the results of the Fishers exact test, performed with the R package TopGO, which examined whether the candidates identified through SNP density analysis as associated with the emergence of DCC3 were enriched with particular Molecular Function GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.16: Summary of the results of the Fishers exact test, performed with the R package TopGO, which examined whether the candidates identified through SNP density analysis as associated with the emergence of DCC3 were enriched with particular Biological function GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.17: Summary of the results of the Fishers exact test, performed with the R package TopGO, which examined whether the candidates identified through SNP density analysis as associated with the emergence of DCC3 were enriched with particular Cellular Component GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.18: Table summarizing the KEGG pathways to which candidates identified through the SNP density analysis were mapped using the program Blast2GO.

Appendix table 1.19: Pangenome gene presence absence table showing the breakdown of the 35,994 genes identified by Roary to make up the MABSC pangenome.

Appendix table 1.20: Functional summary of the four genes encoded by representatives of each of the DCCs and no further isolates in the MABSC global population. The proportion of DCC isolates that encoded these genes suggested that they had been acquired after the clonal expansion of the DCCs and thus were not associated with their initial emergence.

Appendix table 1.21: Emergence of DCC1 - pangnome candidates
Table summarising the 183 genes present in 90% of DCC1 and less than 10% of non-DCC isolates. This includes the Prokka, Pfam and InterPro annotations of each candidate as well the COG functions predicted by EggNOG and *M. tuberculosis* orthologs.

Appendix table 1.22: Emergence of DCC2 - pangnome candidates
Table summarising the 217 genes present in 90% of DCC2 and less than 10% of non-DCC isolates. This includes the Prokka, Pfam and InterPro annotations of each candidate as well the COG functions predicted by EggNOG and *M. tuberculosis* orthologs.

Appendix table 1.23: Emergence of DCC3 - pangnome candidates
Table summarising the 119 genes present in 90% of DCC3 and less than 10% of non-DCC isolates. This includes the Prokka, Pfam and InterPro annotations of each candidate as well the COG functions predicted by EggNOG and *M. tuberculosis* orthologs.

Appendix table 1.24: List of the GO-terms assigned to 4,138 of the 18,386 genes in the MABSC accessory genome by InterProScan.

Appendix table 1.25: Summary of the results of the Fishers exact test, performed with the R package TopGO. This examined whether the candidates identified through pangenome analysis as associated with the emergence of DCC1 were enriched with particular Molecular Function GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.26: Summary of the results of the Fishers exact test, performed with the R package TopGO. This examined whether the candidates identified through pangenome analysis analysis as associated with the emergence of DCC1 were enriched with particular Biological Process function GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.27: Summary of the results of the Fishers exact test, performed with the R package TopGO. This examined whether the candidates identified through pangenome analysis as associated with the emergence of DCC1 were enriched with particular Cellular Component GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.28: Summary of the results of the Fishers exact test, performed with the R package TopGO. This examined whether the candidates identified through pangenome analysis as associated with the emergence of DCC2 were enriched with particular Molecular Function GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.29: Summary of the results of the Fishers exact test, performed with the R package TopGO. This examined whether the candidates identified through pangenome analysis as associated with the emergence of DCC2 were enriched with particular Biological

Process GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.30: Summary of the results of the Fishers exact test, performed with the R package TopGO. This examined whether the candidates identified through pangenome analysis as associated with the emergence of DCC2 were enriched with particular Cellular Component GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.31: Summary of the results of the Fishers exact test, performed with the R package TopGO. This examined whether the candidates identified through pangenome analysis as associated with the emergence of DCC3 were enriched with particular Molecular Function GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.32: Summary of the results of the Fishers exact test, performed with the R package TopGO. This examined whether the candidates identified through pangenome analysis as associated with the emergence of DCC3 were enriched with particular Biological Processes GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.33: Summary of the results of the Fishers exact test, performed with the R package TopGO. This examined whether the candidates identified through pangenome analysis as associated with the emergence of DCC3 were enriched with particular Cellular Component GO-terms. No GO-terms were found to be enriched after the p-values were corrected for multiple testing using the Benjamini-Hochberg method.

Appendix table 1.34: List of the 38 KEGG pathways that pangenome candidates for all the DCCs were predicted to participate in with Blast2GO.

8.2. Appendix for Chapter 3

Appendix table 2.1: Summary of the TreeGubbins clusters identified within the maximum likelihood phylogeny of the *M. a. abscessus* subspecies when all the *M. a. abscessus* isolates were mapped to *M. a. abscessus* ATCC19977.

Appendix table 2.2: Summary of the TreeGubbins clusters identified within i) the maximum likelihood phylogeny of M*. a. bolletii* based on SNP alignment produced after mapping all the *M. a. bolletii* isolates to *M. a. bolletii* BD and II) the clusters identified within the maximum likelihood phylogeny of the *M. a. bolletii* subspecies based on the SNP alignment produced after mapping all the M. a. bolletii isolates to *M. a. abscessus* ATCC19977.

Appendix table 2.3: Summary of the TreeGubbins clusters identified within i) the maximum likelihood phylogeny of M*. a. massiliense* based on SNP alignment produced after mapping all the *M. a. massiliense* isolates to *M. a massiliense* CIP108297 and II) the clusters identified within the maximum likelihood phylogeny of the *M. a. massiliense* subspecies based on the SNP alignment produced after mapping all the *M. a. massiliense* isolates to *M. a. abscessus* ATCC19977.

Appendix table 2.4: Breakdown of all SNPs identified on branches within *M. a. abscessus* clusters when *M. a. abscessus* isolates were mapped to *M. a. abscessus* ATCC19977. This includes SNPs on the terminal branches and those subsequently removed due to recombination.

Appendix table 2.5: Breakdown of all SNPs identified on branches within *M. a. bolletii* clusters when *M. a. bolletii* isolates were mapped to *M. a. bolletii* BD. This includes SNPs on the terminal branches and those subsequently removed due to recombination.

Appendix table 2.6: Breakdown of all SNPs identified on branches within *M. a. massiliense* clusters when *M. a. massiliense* isolates were mapped to *M. a. massiliense* CIP108297. This includes SNPs on the terminal branches and those subsequently removed due to recombination.

Appendix table 2.7: Summary of the binomial test results for the nonsynonymous SNPs accumulated in each gene on branches after the clonal expansion of the *M. a. abscessus* clustered lineages after being mapped to *M. a. abscessus* ATCC19977. Two genes

accumulated a significant number of nonsynonymous SNPs ($p < 0.01$). Terminal branch SNPs and recombination were removed.

Appendix table 2.8: Summary of the binomial test results for the nonsynonymous SNPs accumulated in each gene on branches after the clonal expansion of the *M. a. bolletii* clustered lineages after being mapped to *M. a. bolletii* BD. no genes accumulated a significant number of nonsynonymous SNPs ($p < 0.01$). Terminal branch SNPs and recombination were removed.

Appendix table 2.9:Summary of the binomial test results for the nonsynonymous SNPs accumulated in each gene on branches after the clonal expansion of the *M. a. massiliense* clustered lineages after being mapped to *M. a. massiliense* CIP108297. One gene accumulated a significant number of nonsynonymous SNPs ($p < 0.01$). Terminal branch SNPs and recombination were removed.

Appendix table 2.10: Breakdown of all SNPs identified on branches within the clusters detected by TreeGubbins after all the isolates were mapped to *M. a. abscessus* ATCC19977. This includes SNPs on the terminal branches and those subsequently removed due to recombination.

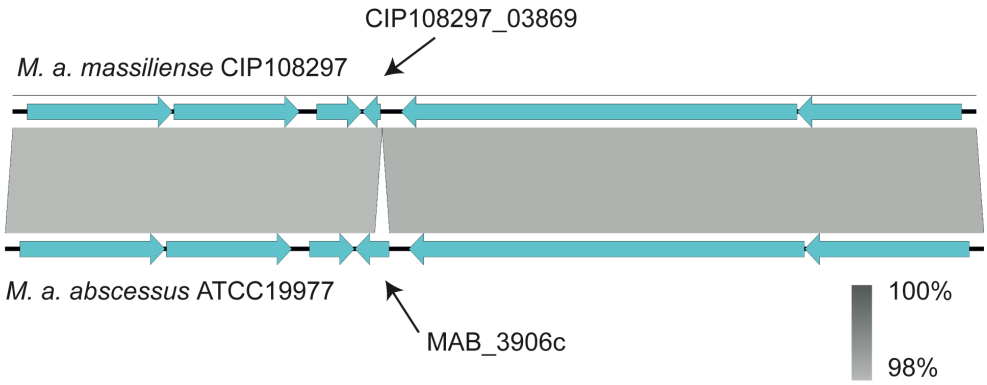
Appendix table 2.11: Summary of the binomial test results based on the SNPs accumulated by each gene on branches after the clonal expansion of all clustered lineages after all isolates were mapped to *M. a. abscessus* ATCC19977. One gene accumulated a significant number of nonsynonymous SNPs ($p < 0.01$). Terminal branch SNPs and recombination were removed.

Appendix table 2.12: Summary of the SNPs removed due to occurring in SNP dense regions after the clonal expansion of *M. a. abscessus* clustered lineages. The regions overlapping known phage or other mobile elements are marked.

Appendix table 2.13: Summary of the SNPs removed due to occurring in SNP dense regions after the clonal expansion of the *M. a. bolletii* clustered lineages detected from the phylogeny inferred from the alignment after the *M. a. bolletii* isolates were mapped to *M. a. bolletii* BD. The regions overlapping known phage or other mobile elements are marked.

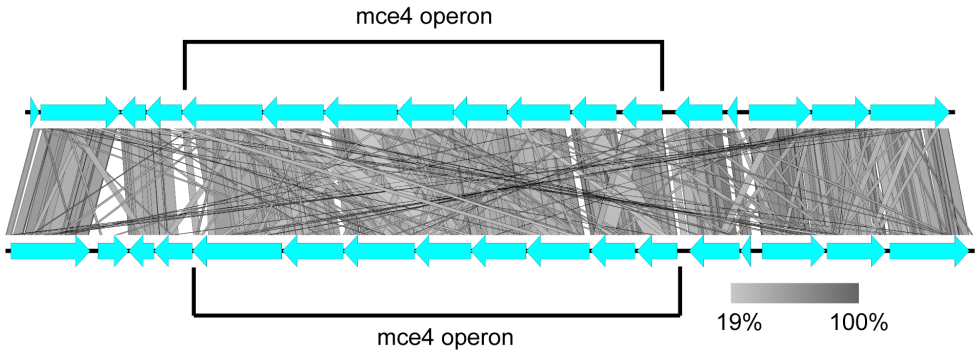
Appendix table 2.14: Summary of the SNPs removed due to occurring in SNP dense regions after the clonal expansion of the *M. a. massiliense* clustered lineages detected from the phylogeny inferred from the alignment after the *M. a. massiliense* isolates were mapped to *M. a. massiliense* CIP108297. The regions overlapping known phage or other mobile elements are marked.

Appendix table 2.15: Summary of the SNPs removed due to occurring in SNP dense regions after the clonal expansion of the clustered lineages detected from the phylogenies inferred from the alignments after all the isolates were mapped to *M. a. abscessus* ATCC19977. The regions overlapping known phage or other mobile elements are marked.



Appendix Figure 2.1: Deletion of the start of CIP108297_03869 in *M. a. massiliense* CIP108297

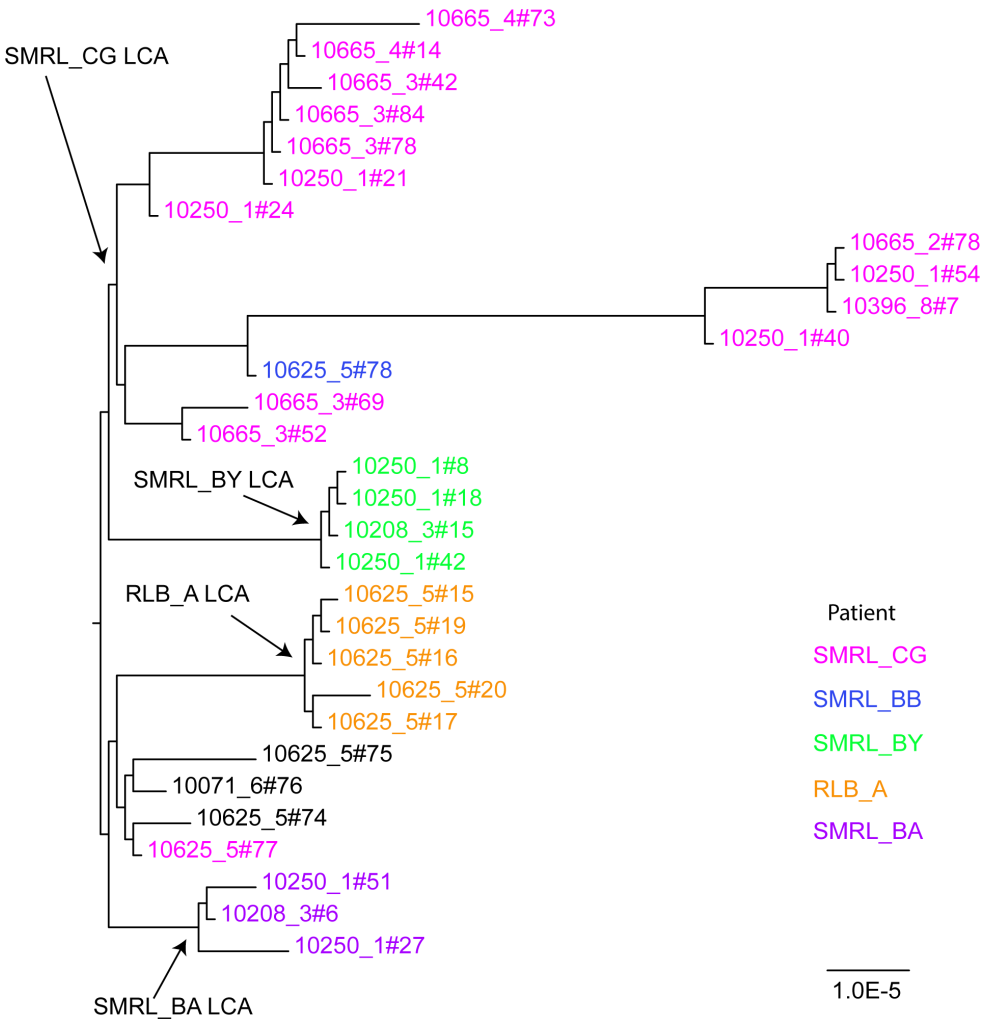
Blastn comparison between the regions in *M. a. massiliense* CIP108297 encoding the candidate gene CIP108297_03869 and the corresponding region in *M. a. abscessus* ATCC19977. This shows that a longer CDS is encoded at this position in *M. a. abscessus* ATCC19977 and suggests that an 111bp deletion has occurred at the start of CIP108297_03869.



Appendix Figure 2.2: Orthology between the *M. tuberculosis mce4* operon and one of the mce operons encoded by *M. a. abscessus* ATCC19977.

tblastx comparison between the *mce4* operon and flanking genes in *M. tuberculosis* H37Rv and an *M. a. abscessus* ATCC19977 shows that the *mce4* operon appears to be ortholgous between the two species, with a high level of sequence conservation and the mce4 operon occurring in the same sequence context. Mce4 appears to be the only orthologous *mce* operon shared between *M. a. abscessus* ATCC19977 and *M. tuberculosis* H37Rv

8.3. Appendix for Chapter 4



Appendix Figure 3.1: Example of the phylogenetic analysis performed to determine the isoaltes to include to examine how the MABSC was adapting to the lung

Phylogenetic analysis to remove patients with polyclonal isolates or isolates that have been transmitted from another patient. 13/14 isolates from patient SMRL_CG form a paraphyletic clade with the only isolate from patient SMRL_BB. The isolate from SMRL_BB is nested within the diversity form patient SMRL_CG, consequently the evolution that occurs after the last common ancestor (LCA) of the 13 SMRL_CG isolates likely occurs under the within host selection pressure, with possible transmission to patient SMRL_BB. The final isolate obtained from patient SMRL_CG is not included in the longitudinal dataset as it does not share a LCA with any other SMRL_CG isolates. Three examples of patients where the isolates form a monophyletic clade, and thus the evolution since the LCA is believed to have occurred under the within host selection pressure, are also marked.

Appendix table 3.1: Summary of the within patient evolution dataset with the reasoning for an isolates inclusion or exclusion. Phylogenetic analysis was used to determine whether all isolates from a patient shared a LCA and thus that the variation observed had occurred within the host or whether an isolate shared a LCA with an isolate from another patient which could suggest the isolate was transmitted and thus it had not evolved under the selection pressure from a single host. Only isolates evolving within a single patient were included in this analysis.

Appendix 3 table 3.2: Summary of all the variants accumulated over time within each patient. The variants were detected by mapping all isolates to the *M. a. abscessus* ATCC19977 reference genome.

Appendix table 3.3: Summary of the number of SNPs accumulated over time within each patient for which the day/month/year of sample collection was known. This dataset was used for the hypermutation analysis.

Appendix table 3.4: The distribution of the 1,185 SNPs accumulated across the *M. a. abscessus* ATCC19977 genome overtime within the 186 MABSC lineages for which the within host evolution was investigated.

Appendix table 3.5: Table showing the results of the binomial test carried out upon each gene to identify if it had accumulated a significant number of nonsynoymous SNPs over time within multiple patients.

Appendix table 3.6: Table reporting the level of differential expression for each of the significantly differentially expressed genes identified from each of the comparisons performed in Table 13. The functional interpretation of each of the differentially expressed genes is reported.

8.4. Appendix for Chapter 5

Appendix table 4.1: Summary of the 189 isolates obtained from nine states in Brazil where post-surgical wound infection outbreaks occurred. One isolate (16933_5#71) was found to be contaminated after sequencing. This dataset was supplemented by two previously

sequenced isolates from Brazil CRM-020 and GO-06 as well as the single isolate per patient MABSC global population dataset (appendix table 5.2).

Appendix table 4.2: Annotation of Del_12078_1#71_01025_01041
The Prokka, InterPro and GO-term annotations of the CDSs present within the globally circulating clone cluster 2 isolate, PAP1174, but lost in the Brazil lineage isolate, BRA_PA_42, starting at position 639781. (This is the start position of the deletion in BRA_PA_42 when the BRA_PA_42 assembly begins with CDS LgrD_9).

Appendix table 4.3: Annotation of Del_12078_1#71_02620_02653
The Prokka, InterPro and GO-term annotations of the CDSs present within the globally circulating clone cluster 2 isolate, PAP1174, but lost in the Brazil lineage isolate, BRA_PA_42, starting at position 2076766 (This is the start position of the deletion in BRA_PA_42 when the BRA_PA_42 assembly begins with CDS LgrD_9).

Appendix table 4.4: Prokka, Pfam and InterPro annotations of the CDs identified on the novel plasmid, pMAB02, associated with the Brazilian epidemic lineage.

8.5. Appendix for Chapter 7

Appendix table 5.1**:** MABSC global population dataset
Summary of the MABSC global population dataset used in this thesis. This includes the date, time, CF status, subspecies, location of isolation of the isolates obtained from each patient . Columns marking which isolates make up the single isolate per patient, within host evolution and pangenome dataset are also included.

Appendix table 5.2**:** This table is a summary of the 29 publicly available MABSC sequences that were added to the MABSC global population dataset sequenced for Bryant et al's global population study. Isolates marked with * represent isolates were the raw reads for the WGS were not available and thus the assembly was shredded using an inhouse script to regenerate the raw reads. ^ marks isolates which were mapped to the *M. a. abscessus* ATCC19977 reference genome with the percent identity to report a mapping at 0.80. If the isolate was known to be associated with an outbreak, the outbreak is recorded in brackets next to the country the isolate was obtained from.

Appendix table 5.3: Summary of the *de novo* assembly statistics for the isolates sequenced in the MABSC global population study.

9. References

1. Navarro S. Historical compilation of cystic fibrosis. Gastroenterología y Hepatología (English Edition). 2016;39(1):36-42.
2. Fanconi G, Uehlinger E, Knauer C. Das Coeliakie-syndrom bei angeborener zystischer Pankreasfibromatose und Bronchiektasien. 86:753-756. Wien Med Wchnschr. 1936;86:753-6.
3. Andersen DH. Cystic Fibrosis of the pancreas and its relation to celiac disease: a clinical and pathologic study. Am J Dis Child. 1938;56(2):344-99.
4. Di Sant'Agnese PEA, Andersen DH. Celiac syndrome; chemotherapy in infections of the respiratory tract associated with cystic fibrosis of the pancreas; observations with penicillin and drugs of the sulfonamide group, with special reference to penicillin aerosol. Am J Dis Child. 1946;72:17-61.
5. Shwachman H, Fekete E, Kulczycki LL, Foley GE. The effect of long-term antibiotic therapy in patients with cystic fibrosis of the pancreas. Antibiot Annu. 1958;6:692-9.
6. Di Sant'Agnese PA, Darling RC, Perera GA, Shea E. Abnormal electrolyte composition of sweat in cystic fibrosis of the pancreas; clinical significance and relationship to the disease. Pediatrics. 1953;12(5):549-63.
7. Andersen DH, Hodges RG. Celiac syndrome; genetics of cystic fibrosis of the pancreas, with a consideration of etiology. Am J Dis Child. 1946;72:62-80.
8. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, et al. Identification of the cystic fibrosis gene: genetic analysis. Science. 1989;245(4922):1073-80.
9. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science. 1989;245(4922):1066-73.
10. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, et al. Identification of the cystic fibrosis gene: chromosome walking and jumping. Science. 1989;245(4922):1059-65.
11. Knowles M, Gatzy J, Boucher R. Increased bioelectric potential difference across respiratory epithelia in cystic fibrosis. N Engl J Med. 1981;305(25):1489-95.
12. Quinton PM. Chloride impermeability in cystic fibrosis. Nature. 1983;301(5899):421-2.
13. Bear CE, Li CH, Kartner N, Bridges RJ, Jensen TJ, Ramjeesingh M, et al. Purification and functional reconstitution of the cystic fibrosis transmembrane conductance regulator (CFTR). Cell. 1992;68(4):809-18.
14. Zhang Z, Chen J. Atomic Structure of the Cystic Fibrosis Transmembrane Conductance Regulator. Cell. 2016;167(6):1586-97.e9.
15. Liu F, Zhang Z, Csanády L, Gadsby DC, Chen J. Molecular Structure of the Human CFTR Ion Channel. Cell. 2017;169(1):85-95.e8.
16. Sosnay PR, Siklosi KR, Van Goor F, Kaniecki K, Yu H, Sharma N, et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. Nat Genet. 2013;45(10):1160-7.
17. De Boeck K, Zolin A, Cuppens H, Olesen HV, Viviani L. The relative frequency of CFTR mutation classes in European patients with cystic fibrosis. J Cyst Fibros. 2014;13(4):403-9.
18. Davies JC, Ebdon A-M, Orchard C. Recent advances in the management of cystic fibrosis. Arch Dis Child. 2014;99(11):1033-6.
19. Bhagirath AY, Li Y, Somayajula D, Dadashi M, Badr S, Duan K. Cystic fibrosis lung environment and *Pseudomonas aeruginosa* infection. BMC Pulm Med. 2016;16(1):174.
20. Pier GB, Grout M, Zaidi T, Meluleni G, Mueschenborn SS, Banting G, et al. Salmonella typhi uses CFTR to enter intestinal epithelial cells. Nature. 1998;393(6680):79-82.

21. Muanprasat C, Chatsudthipong V. Cholera: pathophysiology and emerging therapeutic targets. Future Med Chem. 2013;5(7):781-98.
22. Modiano G, Ciminelli BM, Pignatti PF. Cystic Fibrosis: Cystic fibrosis and lactase persistence: a possible correlation. Eur J Hum Genet. 2006;15:255.
23. Poolman EM, Galvani AP. Evaluating candidate agents of selective pressure for cystic fibrosis. J R Soc Interface. 2007;4(12):91-8.
24. Lubinsky M. Hypothesis: Cystic fibrosis carrier geography reflects interactions of tuberculosis and hypertension with vitamin D deficiency, altitude and temperature. Vitamin D deficiency effects and CF carrier advantage. J Cyst Fibros. 2012;11(1):68-70.
25. Cystic Fibrosis Trust. UK CF Trust annual report 2016 [cited 5 December 2017]. Available from: https://www.cysticfibrosis.org.uk/the-work-we-do/uk-cf-registry/reporting-and-resources.
26. Lopes-Pacheco M. CFTR Modulators: Shedding Light on Precision Medicine for Cystic Fibrosis. Front Pharmacol. 2016;7:275.
27. Wilschanski M, Famini C, Blau H, Rivlin J, Augarten A, Avital A, et al. A pilot study of the effect of gentamicin on nasal potential difference measurements in cystic fibrosis patients carrying stop mutations. Am J Respir Crit Care Med. 2000;161(3 Pt 1):860-5.
28. Kerem E, Hirawat S, Armoni S, Yaakov Y, Shoseyov D, Cohen M, et al. Effectiveness of PTC124 treatment of cystic fibrosis caused by nonsense mutations: a prospective phase II trial. Lancet. 2008;372(9640):719-27.
29. Ramsey BW, Davies J, McElvaney NG, Tullis E, Bell SC, Dřevínek P, et al. A CFTR Potentiator in Patients with Cystic Fibrosis and the G551D Mutation. N Engl J Med. 2011;365(18):1663-72.
30. Dodge JA, Lewis PA, Stanton M, Wilsher J. Cystic fibrosis mortality and survival in the UK: 1947-2003. Eur Respir J. 2007;29(3):522-6.
31. Hauser AR, Jain M, Bar-Meir M, McColley SA. Clinical significance of microbial infection and adaptation in cystic fibrosis. Clin Microbiol Rev. 2011;24(1):29-70.
32. Tang AC, Turvey SE, Alves MP, Regamey N, Tümmler B, Hartl D. Current concepts: host-pathogen interactions in cystic fibrosis airways disease. Eur Respir Rev. 2014;23(133):320-32.
33. Cohen TS, Prince A. Cystic fibrosis: a mucosal immunodeficiency syndrome. Nat Med. 2012;18(4):509-19.
34. Filkins LM, O'Toole GA. Cystic Fibrosis Lung Infections: Polymicrobial, Complex, and Hard to Treat. PLoS Pathog. 2015;11(12):e1005258.
35. Mahenthiralingam E. Emerging cystic fibrosis pathogens and the microbiome. Paediatr Respir Rev. 2014;15 Suppl 1:13-5.
36. Salsgiver EL, Fink AK, Knapp EA, LiPuma JJ, Olivier KN, Marshall BC, et al. Changing Epidemiology of the Respiratory Bacteriology of Patients With Cystic Fibrosis. Chest. 2016;149(2):390-400.
37. Lipuma JJ. The changing microbial epidemiology in cystic fibrosis. Clin Microbiol Rev. 2010;23(2):299-323.
38. Parkins MD, Floto RA. Emerging bacterial pathogens and changing concepts of bacterial pathogenesis in cystic fibrosis. J Cyst Fibros. 2015;14(3):293-304.
39. Marvig RL, Johansen HK, Molin S, Jelsbak L. Genome analysis of a transmissible lineage of *Pseudomonas aeruginosa* reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. PLoS Genet. 2013;9(9):e1003741.
40. Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Jr., et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet. 2011;43(12):1275-80.
41. Winstanley C, O'Brien S, Brockhurst MA. *Pseudomonas aeruginosa* Evolutionary Adaptation and Diversification in Cystic Fibrosis Chronic Lung Infections. Trends Microbiol. 2016;24(5):327-37.

42. Behrends V, Ryall B, Zlosnik JEA, Speert DP, Bundy JG, Williams HD. Metabolic adaptations of *Pseudomonas aeruginosa* during cystic fibrosis chronic lung infections. Environ Microbiol. 2013;15(2):398-408.
43. Kelly N, Tempany E, Falkiner FR, Fitzgerald MX, O'Boyle C, Keane CT. Does *Pseudomonas* Cross-infection occur between Cystic Fibrosis Patients? Lancet. 1982;320(8300):688-90.
44. Grothues D, Koopmann U, von der Hardt H, Tümmler B. Genome fingerprinting of *Pseudomonas aeruginosa* indicates colonization of cystic fibrosis siblings with closely related strains. J Clin Microbiol. 1988;26(10):1973-7.
45. Sun L, Jiang RZ, Steinbach S, Holmes A, Campanelli C, Forstner J, et al. The emergence of a highly transmissible lineage of cbl+ *Pseudomonas* (Burkholderia) cepacia causing CF centre epidemics in North America and Britain. Nat Med. 1995;1(7):661-6.
46. Cheng K, Smyth RL, Govan JR, Doherty C, Winstanley C, Denning N, et al. Spread of beta-lactam-resistant *Pseudomonas aeruginosa* in a cystic fibrosis clinic. Lancet. 1996;348(9028):639-42.
47. Armstrong DS, Nixon GM, Carzino R, Bigham A, Carlin JB, Robins-Browne RM, et al. Detection of a widespread clone of *Pseudomonas aeruginosa* in a pediatric cystic fibrosis clinic. Am J Respir Crit Care Med. 2002;166(7):983-7.
48. Fothergill JL, Walshaw MJ, Winstanley C. Transmissible strains of *Pseudomonas aeruginosa* in cystic fibrosis lung infections. Eur Respir J. 2012;40(1):227-38.
49. Armstrong D, Bell S, Robinson M, Bye P, Rose B, Harbour C, et al. Evidence for spread of a clonal strain of *Pseudomonas aeruginosa* among cystic fibrosis clinics. J Clin Microbiol. 2003;41(5):2266-7.
50. Biddick R, Spilker T, Martin A, LiPuma JJ. Evidence of transmission of Burkholderia cepacia, Burkholderia multivorans and Burkholderia dolosa among persons with cystic fibrosis. FEMS Microbiol Lett. 2003;228(1):57-62.
51. Al-Aloul M, Crawley J, Winstanley C, Hart CA, Ledson MJ, Walshaw MJ. Increased morbidity associated with chronic infection by an epidemic *Pseudomonas aeruginosa* strain in CF patients. Thorax. 2004;59(4):334-6.
52. Renna M, Schaffner C, Brown K, Shang S, Tamayo MH, Hegyi K, et al. Azithromycin blocks autophagy and may predispose cystic fibrosis patients to mycobacterial infection. J Clin Invest. 2011;121(9):3554-63.
53. Olivier KN, Weber DJ, Wallace RJ, Jr., Faiz AR, Lee J-H, Zhang Y, et al. Nontuberculous mycobacteria. I: multicenter prevalence study in cystic fibrosis. Am J Respir Crit Care Med. 2003;167(6):828-34.
54. Rodman DM, Polis JM, Heltshe SL, Sontag MK, Chacon C, Rodman RV, et al. Late diagnosis defines a unique population of long-term survivors of cystic fibrosis. Am J Respir Crit Care Med. 2005;171(6):621-6.
55. Jarand J, Levin A, Zhang L, Huitt G, Mitchell JD, Daley CL. Clinical and microbiologic outcomes in patients receiving treatment for *Mycobacterium abscessus* pulmonary disease. Clin Infect Dis. 2011;52(5):565-71.
56. Magee JG, Ward AC. Mycobacterium. Bergey's Manual of Systematics of Archaea and Bacteria: John Wiley & Sons, Ltd; 2015.
57. Devulder G, Pérouse de Montclos M, Flandrois JP. A multigene approach to phylogenetic analysis using the genus Mycobacterium as a model. Int J Syst Evol Microbiol. 2005;55(Pt 1):293-302.
58. Tortoli E, Fedrizzi T, Meehan CJ, Trovato A, Grottola A, Giacobazzi E, et al. The new phylogeny of the genus Mycobacterium: The old and the news. Infect Genet Evol. 2017;56:19-25.
59. Tortoli E, Kohl TA, Brown-Elliott BA, Trovato A, Leão SC, Garcia MJ, et al. Emended description of *Mycobacterium abscessus*, *Mycobacterium abscessus* subsp. *abscessus* and *Mycobacterium abscessus* subsp. *bolletii* and designation of *Mycobacterium abscessus* subsp. *massiliense* comb. nov. Int J Syst Evol Microbiol. 2016;66(11):4471-9.

60. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5):1792-7.
61. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.
62. Moore M, Frerichs JB. An unusual acid-fast infection of the knee with subcutaneous, abscess-like lesions of the gluteal region; report of a case with a study of the organism, *Mycobacterium abscessus*, n. sp. J Invest Dermatol. 1953;20(2):133-69.
63. Kubica GP, Baess I, Gordon RE, Jenkins PA, Kwapinski JB, McDurmont C, et al. A co-operative numerical analysis of rapidly growing mycobacteria. J Gen Microbiol. 1972;73(1):55-70.
64. Kusunoki S, Ezaki T. 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. Int J Syst Bacteriol. 1992;42(2):240-5.
65. Adékambi T, Reynaud-Gaubert M, Greub G, Gevaudan M-J, La Scola B, Raoult D, et al. Amoebal coculture of "*Mycobacterium massiliense*" sp. nov. from the sputum of a patient with hemoptoic pneumonia. J Clin Microbiol. 2004;42(12):5493-501.
66. Adékambi T, Berger P, Raoult D, Drancourt M. *rpoB* gene sequence-based characterization of emerging non-tuberculous mycobacteria with descriptions of *Mycobacterium bolletii* sp. nov., *Mycobacterium phocaicum* sp. nov. and *Mycobacterium aubagnense* sp. nov. Int J Syst Evol Microbiol. 2006;56(Pt 1):133-43.
67. Euzby J. List of new names and new combinations previously effectively, but not validly, published. Int J Syst Evol Microbiol. 2006;56(9):2025-7.
68. Leão SC, Tortoli E, Viana-Niero C, Ueki SYM, Lima KVB, Lopes ML, et al. Characterization of mycobacteria from a major Brazilian outbreak suggests that revision of the taxonomic status of members of the *Mycobacterium chelonae-M. abscessus* group is needed. J Clin Microbiol. 2009;47(9):2691-8.
69. Leão SC, Tortoli E, Euzéby JP, Garcia MJ. 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*. Int J Syst Evol Microbiol. 2011;61(Pt 9):2311-3.
70. Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, et al. Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. Lancet. 2013;381(9877):1551-60.
71. Cho Y-J, Yi H, Chun J, Cho S-N, Daley CL, Koh W-J, et al. The Genome Sequence of '*Mycobacterium massiliense*' Strain CIP 108297 Suggests the Independent Taxonomic Status of the *Mycobacterium abscessus* Complex at the Subspecies Level. PLoS One. 2013;8(11):e81560.
72. Sassi M, Drancourt M. Genome analysis reveals three genomospecies in *Mycobacterium abscessus*. BMC Genomics. 2014;15(1):359.
73. Bryant JM, Grogono DM, Rodriguez-Rincon D, Everall I, Brown KP, Moreno P, et al. Emergence and spread of a human-transmissible multidrug-resistant nontuberculous mycobacterium. Science. 2016;354(6313):751-7.
74. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007;57(Pt 1):81-91.
75. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106(45):19126-31.
76. Kim M, Oh H-S, Park S-C, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol. 2014;64(Pt 2):346-51.
77. Nash KA, Brown-Elliott BA, Wallace RJ, Jr. A novel gene, *erm(41)*, confers inducible macrolide resistance to clinical isolates of *Mycobacterium abscessus* but is absent from *Mycobacterium chelonae*. Antimicrob Agents Chemother. 2009;53(4):1367-76.

78. Adekambi T, Sassi M, van Ingen J, Drancourt M. Reinstating *Mycobacterium massiliense* and *Mycobacterium bolletii* as species of the *Mycobacterium abscessus* complex. Int J Syst Evol Microbiol. 2017;67(8):2726-30.
79. Macheras E, Konjek J, Roux A-L, Thiberge J-M, Bastian S, Leão SC, et al. Multilocus sequence typing scheme for the *Mycobacterium abscessus* complex. Res Microbiol. 2014;165(2):82-90.
80. Panagea T, Pincus DH, Grogono D, Jones M, Bryant J, Parkhill J, et al. *Mycobacterium abscessus* Complex Identification with Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. J Clin Microbiol. 2015;53(7):2355-8.
81. Choi G-E, Cho Y-J, Koh W-J, Chun J, Cho S-N, Shin SJ. Draft genome sequence of *Mycobacterium abscessus* subsp. *bolletii* BD(T). J Bacteriol. 2012;194(10):2756-7.
82. Ripoll F, Pasek S, Schenowitz C, Dossat C, Barbe V, Rottman M, et al. Non mycobacterial virulence genes in the genome of the emerging pathogen *Mycobacterium abscessus*. PLoS One. 2009;4(6):e5660.
83. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998;393(6685):537-44.
84. Choo SW, Wee WY, Ngeow YF, Mitchell W, Tan JL, Wong GJ, et al. Genomic reconnaissance of clinical isolates of emerging human pathogen *Mycobacterium abscessus* reveals high evolutionary potential. Sci Rep. 2014;4:4061.
85. Howard ST, Byrd TF, Lyons CR. A polymorphic region in *Mycobacterium abscessus* contains a novel insertion sequence element. Microbiology. 2002;148(Pt 10):2987-96.
86. Sapriel G, Konjek J, Orgeur M, Bouri L, Frézal L, Roux A-L, et al. Genome-wide mosaicism within *Mycobacterium abscessus*: evolutionary and epidemiological implications. BMC Genomics. 2016;17:118.
87. Leão SC, Matsumoto CK, Carneiro A, Ramos RT, Nogueira CL, Lima JD, Jr., et al. The detection and sequencing of a broad-host-range conjugative IncP-1β plasmid in an epidemic strain of *Mycobacterium abscessus* subsp. *bolletii*. PLoS One. 2013;8(4):e60746.
88. Sassi M, Gouret P, Chabrol O, Pontarotti P, Drancourt M. Mycobacteriophage-drived diversification of *Mycobacterium abscessus*. Biol Direct. 2014;9:19.
89. Davidson RM, Hasan NA, de Moura VCN, Duarte RS, Jackson M, Strong M. Phylogenomics of Brazilian epidemic isolates of *Mycobacterium abscessus* subsp. *bolletii* reveals relationships of global outbreak strains. Infect Genet Evol. 2013;20:292-7.
90. Tan JL, Ng KP, Ong CS, Ngeow YF. Genomic Comparisons Reveal Microevolutionary Differences in *Mycobacterium abscessus* Subspecies. Front Microbiol. 2017;8:2042.
91. Tettelin H, Davidson RM, Agrawal S, Aitken ML, Shallom S, Hasan NA, et al. High-level relatedness among *Mycobacterium abscessus* subsp. *massiliense* strains from widely separated outbreaks. Emerg Infect Dis. 2014;20(3):364-71.
92. van Ingen J, Boeree MJ, Dekhuijzen PNR, van Soolingen D. Environmental sources of rapid growing nontuberculous mycobacteria causing disease in humans. Clin Microbiol Infect. 2009;15(10):888-93.
93. Halstrom S, Price P, Thomson R. Review: Environmental mycobacteria as a cause of human infection. Int J Mycobacteriol. 2015;4(2):81-91.
94. De Groote MA, Pace NR, Fulton K, Falkinham JO, 3rd. Relationships between Mycobacterium isolates from patients with pulmonary mycobacterial infection and potting soils. Appl Environ Microbiol. 2006;72(12):7602-6.
95. Williams MM, Yakrus MA, Arduino MJ, Cooksey RC, Crane CB, Banerjee SN, et al. Structural analysis of biofilm formation by rapidly and slowly growing nontuberculous mycobacteria. Appl Environ Microbiol. 2009;75(7):2091-8.
96. Brown-Elliott BA, Wallace RJ, Jr., Tichindelean C, Sarria JC, McNulty S, Vasireddy R, et al. Five-year outbreak of community- and hospital-acquired *Mycobacterium porcinum* infections related to public water supplies. J Clin Microbiol. 2011;49(12):4231-8.

97. Carson LA, Petersen NJ, Favero MS, Aguero SM. Growth characteristics of atypical mycobacteria in water and their comparative resistance to disinfectants. Appl Environ Microbiol. 1978;36(6):839-46.
98. Zhang Y, Rajagopalan M, Brown BA, Wallace RJ, Jr. Randomly amplified polymorphic DNA PCR for comparison of *Mycobacterium abscessus* strains from nosocomial outbreaks. J Clin Microbiol. 1997;35(12):3132-9.
99. Appelgren P, Farnebo F, Dotevall L, Studahl M, Jönsson B, Petrini B. Late-onset posttraumatic skin and soft-tissue infections caused by rapid-growing mycobacteria in tsunami survivors. Clin Infect Dis. 2008;47(2):e11-6.
100. Thomson R, Tolson C, Carter R, Coulter C, Huygens F, Hargreaves M. Isolation of nontuberculous mycobacteria (NTM) from household water and shower aerosols in patients with pulmonary disease caused by NTM. J Clin Microbiol. 2013;51(9):3006-11.
101. Thomson R, Tolson C, Sidjabat H, Huygens F, Hargreaves M. *Mycobacterium abscessus* isolated from municipal water - a potential source of human infection. BMC Infect Dis. 2013;13:241.
102. Gao LY, Harb OS, Abu Kwaik Y. Utilization of similar mechanisms by *Legionella pneumophila* to parasitize two evolutionarily distant host cells, mammalian macrophages and protozoa. Infect Immun. 1997;65(11):4738-46.
103. September SM, Brözel VS, Venter SN. Diversity of nontuberculoid Mycobacterium species in biofilms of urban and semiurban drinking water distribution systems. Appl Environ Microbiol. 2004;70(12):7571-3.
104. Fennelly KP, Ojano-Dirain C, Yang Q, Liu L, Lu L, Progulske-Fox A, et al. Biofilm Formation by *Mycobacterium abscessus* in a Lung Cavity. Am J Respir Crit Care Med. 2016;193(6):692-3.
105. Smith MJ, Efthimiou J, Hodson ME, Batten JC. Mycobacterial isolations in young adults with cystic fibrosis. Thorax. 1984;39(5):369-75.
106. Hjelte L, Petrini B, Källenius G, Strandvik B. Prospective study of mycobacterial infections in patients with cystic fibrosis. Thorax. 1990;45(5):397-400.
107. Roux A-L, Catherinot E, Ripoll F, Soismier N, Macheras E, Ravilly S, et al. Multicenter study of prevalence of nontuberculous mycobacteria in patients with cystic fibrosis in france. J Clin Microbiol. 2009;47(12):4124-8.
108. Esther CR, Jr., Esserman DA, Gilligan P, Kerr A, Noone PG. Chronic *Mycobacterium abscessus* infection and lung function decline in cystic fibrosis. J Cyst Fibros. 2010;9(2):117-23.
109. Catherinot E, Roux A-L, Vibet M-A, Bellis G, Ravilly S, Lemonnier L, et al. *Mycobacterium avium* and *Mycobacterium abscessus* complex target distinct cystic fibrosis patient subpopulations. J Cyst Fibros. 2013;12(1):74-80.
110. Adjemian J, Olivier KN, Prevots DR. Nontuberculous mycobacteria among patients with cystic fibrosis in the United States: screening practices and environmental risk. Am J Respir Crit Care Med. 2014;190(5):581-6.
111. Sermet-Gaudelus I, Le Bourgeois M, Pierre-Audigier C, Offredo C, Guillemot D, Halley S, et al. *Mycobacterium abscessus* and children with cystic fibrosis. Emerg Infect Dis. 2003;9(12):1587-91.
112. Levy I, Grisaru-Soen G, Lerner-Geva L, others. Multicenter cross-sectional study of nontuberculous mycobacterial infections among cystic fibrosis patients, Israel. Emerg Infect Dis. 2008.
113. Floto RA, Olivier KN, Saiman L, Daley CL, Herrmann J-L, Nick JA, et al. US Cystic Fibrosis Foundation and European Cystic Fibrosis Society consensus recommendations for the management of non-tuberculous mycobacteria in individuals with cystic fibrosis. Thorax. 2016;71 Suppl 1:i1-22.
114. Koh W-J, Chang B, Jeong B-H, Jeon K, Kim S-Y, Lee NY, et al. Increasing Recovery of Nontuberculous Mycobacteria from Respiratory Specimens over a 10-Year Period in a Tertiary Referral Hospital in South Korea. Tuberc Respir Dis. 2013;75(5):199-204.

115. Simons S, van Ingen J, Hsueh P-R, Van Hung N, Dekhuijzen PNR, Boeree MJ, et al. Nontuberculous mycobacteria in respiratory tract infections, eastern Asia. Emerg Infect Dis. 2011;17(3):343-9.
116. Thomson RM, Others. Changing epidemiology of pulmonary nontuberculous mycobacteria infections. Emerg Infect Dis. 2010;16(10):1576.
117. Feazel LM, Baumgartner LK, Peterson KL, Frank DN, Harris JK, Pace NR. Opportunistic pathogens enriched in showerhead biofilms. Proc Natl Acad Sci U S A. 2009;106(38):16393-9.
118. Falkinham JO, 3rd. Nontuberculous mycobacteria from household plumbing of patients with nontuberculous mycobacteria disease. Emerg Infect Dis. 2011;17(3):419-24.
119. Aitken ML, Limaye A, Pottinger P, Whimbey E, Goss CH, Tonelli MR, et al. Respiratory Outbreak of *Mycobacterium abscessus* Subspecies *massiliense* in a Lung Transplant and Cystic Fibrosis Center. Am J Respir Crit Care Med. 2012;185(2):231-2.
120. Lee M-R, Sheng W-H, Hung C-C, Yu C-J, Lee L-N, Hsueh P-R. *Mycobacterium abscessus* Complex Infections in Humans. Emerg Infect Dis. 2015;21(9):1638-46.
121. Lee M-R, Cheng A, Lee Y-C, Yang C-Y, Lai C-C, Huang Y-T, et al. CNS infections caused by *Mycobacterium abscessus* complex: clinical features and antimicrobial susceptibilities of isolates. J Antimicrob Chemother. 2012;67(1):222-5.
122. Lee M-R, Ko J-C, Liang S-K, Lee S-W, Yen DH-T, Hsueh P-R. Bacteraemia caused by *Mycobacterium abscessus* subsp. *abscessus* and *M. abscessus* subsp. *bolletii*: clinical features and susceptibilities of the isolates. Int J Antimicrob Agents. 2014;43(5):438-41.
123. van Ingen J, Looijmans F, Mirck P, Dekhuijzen R, Boeree M, van Soolingen D. Otomastoiditis caused by *Mycobacterium abscessus*, The Netherlands. Emerg Infect Dis. 2010;16(1):166-8.
124. Umrao J, Singh D, Zia A, Saxena S, Sarsaiya S, Singh S, et al. Prevalence and species spectrum of both pulmonary and extrapulmonary nontuberculous mycobacteria isolates at a tertiary care center. Int J Mycobacteriol. 2016;5(3):288-93.
125. Kham-ngam I, Chetchotisakd P, Ananta P, Chaimanee P, Sadee P, Reechaipichitkul W, et al. Epidemiology of and risk factors for extrapulmonary nontuberculous mycobacterial infections in Northeast Thailand. PeerJ. 2018;6:e5479.
126. Stout JE, Gadkowski LB, Rath S, Alspaugh JA, Miller MB, Cox GM. Pedicure-associated rapidly growing mycobacterial infection: an endemic disease. Clin Infect Dis. 2011;53(8):787-92.
127. Wu C-H, Thong H-Y, Huang C-C, Chen P-H. Report of two cases of cutaneous *Mycobacterium abscessus* infection complicating professional decorative tattoo. Dermatologica Sinica. 2017;35(1):40-3.
128. Tang P, Walsh S, Murray C, Alterman C, Varia M, Broukhanski G, et al. Outbreak of acupuncture-associated cutaneous *Mycobacterium abscessus* infections. J Cutan Med Surg. 2006;10(4):166-9.
129. David S, Douglas HE, Joanna G, Alison R, Barry MA, Katherine AF, et al. Multistate US Outbreak of Rapidly Growing Mycobacterial Infections Associated with Medical Tourism to the Dominican Republic, 2013–20141. Emerging Infectious Disease journal. 2016;22(8):1340.
130. Leão SC, Viana-Niero C, Matsumoto CK, Lima KVB, Lopes ML, Palaci M, et al. Epidemic of surgical-site infections by a single clone of rapidly growing mycobacteria in Brazil. Future Microbiol. 2010;5(6):971-80.
131. Lorena NSdO, Pitombo MB, Côrtes PB, Maya MCA, Silva MGd, Carvalho ACdS, et al. Mycobacterium *massiliense* BRA100 strain recovered from postsurgical infections: resistance to high concentrations of glutaraldehyde and alternative solutions for high level disinfection. Acta Cir Bras. 2010;25(5):455-9.
132. Griffith DE, Aksamit T, Brown-Elliott BA, Catanzaro A, Daley C, Gordin F, et al. An official ATS/IDSA statement: diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. Am J Respir Crit Care Med. 2007;175(4):367-416.

133. Griffith DE. The talking *Mycobacterium abscessus* blues. Clin Infect Dis. 2011;52(5):572-4.
134. Ryu YJ, Koh W-J, Daley CL. Diagnosis and Treatment of Nontuberculous Mycobacterial Lung Disease: Clinicians' Perspectives. Tuberc Respir Dis. 2016;79(2):74-84.
135. Griffith DE, Girard WM, Wallace RJ, Jr. Clinical features of pulmonary disease caused by rapidly growing mycobacteria. Am Rev Respir Dis. 1993.
136. Jeon K, Kwon OJ, Lee NY, Kim B-J, Kook Y-H, Lee S-H, et al. Antibiotic treatment of *Mycobacterium abscessus* lung disease: a retrospective analysis of 65 patients. Am J Respir Crit Care Med. 2009;180(9):896-902.
137. Lyu J, Jang HJ, Song JW, Choi C-M, Oh Y-M, Lee SD, et al. Outcomes in patients with *Mycobacterium abscessus* pulmonary disease treated with long-term injectable drugs. Respir Med. 2011;105(5):781-7.
138. Gilljam M, Scherstén H, Silverborn M, Jönsson B, Ericsson Hollsing A. Lung transplantation in patients with cystic fibrosis and *Mycobacterium abscessus* infection. J Cyst Fibros. 2010;9(4):272-6.
139. Taylor JL, Palmer SM. *Mycobacterium abscessus* chest wall and pulmonary infection in a cystic fibrosis lung transplant recipient. J Heart Lung Transplant. 2006;25(8):985-8.
140. Tissot A, Thomas MF, Corris PA, Brodlie M. NonTuberculous Mycobacteria infection and lung transplantation in cystic fibrosis: a worldwide survey of clinical practice. BMC Pulm Med. 2018;18(1):86.
141. Koh WJ, Stout JE, Yew WW. Advances in the management of pulmonary disease due to *Mycobacterium abscessus* complex. Int J Tuberc Lung Dis. 2014;18(10):1141-8.
142. Mougari F, Bouziane F, Crockett F, Nessar R, Chau F, Veziris N, et al. Selection of Resistance to Clarithromycin in *Mycobacterium abscessus* Subspecies. Antimicrob Agents Chemother. 2017;61(1).
143. Bastian S, Veziris N, Roux A-L, Brossier F, Gaillard J-L, Jarlier V, et al. Assessment of clarithromycin susceptibility in strains belonging to the *Mycobacterium abscessus* group by erm(41) and rrl sequencing. Antimicrob Agents Chemother. 2011;55(2):775-81.
144. Yoshida S, Tsuyuguchi K, Suzuki K, Tomita M, Okada M, Hayashi S, et al. Further isolation of *Mycobacterium abscessus* subsp. *abscessus* and subsp. *bolletii* in different regions of Japan and susceptibility of these isolates to antimicrobial agents. Int J Antimicrob Agents. 2013;42(3):226-31.
145. Brown-Elliott BA, Vasireddy S, Vasireddy R, Iakhiaeva E, Howard ST, Nash K, et al. Utility of sequencing the erm(41) gene in isolates of *Mycobacterium abscessus* subsp. *abscessus* with low and intermediate clarithromycin MICs. J Clin Microbiol. 2015;53(4):1211-5.
146. Shallom SJ, Moura NS, Olivier KN, Sampaio EP, Holland SM, Zelazny AM. New Real-Time PCR Assays for Detection of Inducible and Acquired Clarithromycin Resistance in the *Mycobacterium abscessus* Group. J Clin Microbiol. 2015;53(11):3430-7.
147. Tung Y-J, Bittaye SO, Tsai J-R, Lin C-Y, Huang C-H, Chen T-C, et al. Risk factors for microbiologic failure among Taiwanese adults with *Mycobacterium abscessus* complex pulmonary disease. J Microbiol Immunol Infect. 2015;48(4):437-45.
148. van Ingen J, Wagner D, Gallagher J, Morimoto K, Lange C, Haworth CS, et al. Poor adherence to management guidelines in nontuberculous mycobacterial pulmonary diseases. Eur Respir J. 2017;49(2).
149. Nessar R, Cambau E, Reyrat JM, Murray A, Gicquel B. *Mycobacterium abscessus*: a new antibiotic nightmare. J Antimicrob Chemother. 2012;67(4):810-8.
150. Poehlsgaard J, Douthwaite S. Macrolide antibiotic interaction and resistance on the bacterial ribosome. Curr Opin Investig Drugs. 2003;4(2):140-8.
151. Morris RP, Nguyen L, Gatfield J, Visconti K, Nguyen K, Schnappinger D, et al. Ancestral antibiotic resistance in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A. 2005;102(34):12200-5.

152. Hurst-Hess K, Rudra P, Ghosh P. *Mycobacterium abscessus* WhiB7 Regulates a Species-Specific Repertoire of Genes To Confer Extreme Antibiotic Resistance. Antimicrob Agents Chemother. 2017;61(11).
153. Sreevatsan S, Stockbauer KE, Pan X, Kreiswirth BN, Moghazeh SL, Jacobs WR, Jr., et al. Ethambutol resistance in *Mycobacterium tuberculosis*: critical role of embB mutations. Antimicrob Agents Chemother. 1997;41(8):1677-81.
154. Matrat S, Aubry A, Mayer C, Jarlier V, others. Mutagenesis in the α3α4 GyrA Helix and in the Toprim Domain of GyrB Refines the Contribution of Mycobacterium tuberculosis DNA Gyrase to Intrinsic Resistance to Quinolones. Antimicrob Agents Chemother. 2008.
155. Wallace RJ, Jr., Meier A, Brown BA, Zhang Y, Sander P, Onyi GO, et al. Genetic basis for clarithromycin resistance among isolates of *Mycobacterium chelonae* and *Mycobacterium abscessus*. Antimicrob Agents Chemother. 1996;40(7):1676-81.
156. Nessar R, Reyrat JM, Murray A, Gicquel B. Genetic analysis of new 16S rRNA mutations conferring aminoglycoside resistance in *Mycobacterium abscessus*. J Antimicrob Chemother. 2011;66(8):1719-24.
157. Stout JE, Koh W-J, Yew WW. Update on pulmonary disease due to non-tuberculous mycobacteria. Int J Infect Dis. 2016;45:123-34.
158. Bernut A, Herrmann J-L, Kissa K, Dubremetz J-F, Gaillard J-L, Lutfalla G, et al. *Mycobacterium abscessus* cording prevents phagocytosis and promotes abscess formation. Proc Natl Acad Sci U S A. 2014;111(10):E943-52.
159. Bonaiti G, Pesci A, Marruchella A, Lapadula G, Gori A, Aliberti S. Nontuberculous Mycobacteria in Noncystic Fibrosis Bronchiectasis. Biomed Res Int. 2015;2015:197950.
160. Roux A-L, Viljoen A, Bah A, Simeone R, Bernut A, Laencina L, et al. The distinct fate of smooth and rough *Mycobacterium abscessus* variants inside macrophages. Open Biol. 2016;6(11).
161. Ripoll F, Deshayes C, Pasek S, Laval F, Beretti J-L, Biet F, et al. Genomics of glycopeptidolipid biosynthesis in *Mycobacterium abscessus* and M. chelonae. BMC Genomics. 2007;8:114.
162. Pawlik A, Garnier G, Orgeur M, Tong P, Lohan A, Le Chevalier F, et al. Identification and characterization of the genetic changes responsible for the characteristic smooth-to-rough morphotype alterations of clinically persistent *Mycobacterium abscessus*. Mol Microbiol. 2013;90(3):612-29.
163. Catherinot E, Clarissou J, Etienne G, Ripoll F, Emile JF, Daffé M, et al. Hypervirulence of a Rough Variant of the *Mycobacterium abscessus* Type Strain. Infect Immun. 2007;75(2):1055-8.
164. Rhoades ER, Archambault AS, Greendyke R, Hsu F-F, Streeter C, Byrd TF. *Mycobacterium abscessus* glycopeptidolipids mask underlying cell wall phosphatidyl-myo-inositol mannosides blocking induction of human macrophage TNFα by preventing interaction with TLR2. The Journal of Immunology. 2009:jimmunol-0802181.
165. Roux A-L, Ray A, Pawlik A, Medjahed H, Etienne G, Rottman M, et al. Overexpression of proinflammatory TLR-2-signalling lipoproteins in hypervirulent mycobacterial variants. Cell Microbiol. 2011;13(5):692-704.
166. Nessar R, Reyrat J-M, Davidson LB, Byrd TF. Deletion of the mmpL4b gene in the *Mycobacterium abscessus* glycopeptidolipid biosynthetic pathway results in loss of surface colonization capability, but enhanced ability to replicate in human macrophages and stimulate their innate immune response. Microbiology. 2011;157(Pt 4):1187-95.
167. Byrd TF, Lyons CR. Preliminary characterization of a *Mycobacterium abscessus* mutant in human and murine models of infection. Infect Immun. 1999;67(9):4700-7.
168. Howard ST, Rhoades E, Recht J, Pang X, Alsup A, Kolter R, et al. Spontaneous reversion of *Mycobacterium abscessus* from a smooth to a rough morphotype is associated with reduced expression of glycopeptidolipid and reacquisition of an invasive phenotype. Microbiology. 2006;152(Pt 6):1581-90.
169. Medjahed H, Gaillard J-L, Reyrat J-M. *Mycobacterium abscessus*: a new player in the mycobacterial field. Trends Microbiol. 2010;18(3):117-23.

170. Grzegorzewicz AE, de Sousa-d'Auria C, McNeil MR, Huc-Claustre E, Jones V, Petit C, et al. Assembling of the *Mycobacterium tuberculosis* Cell Wall Core. J Biol Chem. 2016;291(36):18867-79.
171. Laencina L, Dubois V, Le Moigne V, Viljoen A, Majlessi L, Pritchard J, et al. Identification of genes required for *Mycobacterium abscessus* growth in vivo with a prominent role of the ESX-4 locus. Proc Natl Acad Sci U S A. 2018;115(5):E1002-E11.
172. Halloum I, Carrère-Kremer S, Blaise M, Viljoen A, Bernut A, Le Moigne V, et al. Deletion of a dehydratase important for intracellular growth and cording renders rough *Mycobacterium abscessus* avirulent. Proc Natl Acad Sci U S A. 2016;113(29):E4228-37.
173. Kim YS, Yang C-S, Nguyen LT, Kim JK, Jin HS, Choe Jh, et al. *Mycobacterium abscessus* ESX-3 plays an important role in host inflammatory and pathological responses during infection. Microbes Infect. 2017;19(1):5-17.
174. Akhter Y, Ehebauer MT, Mukhopadhyay S, Hasnain SE. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: perhaps more? Biochimie. 2012;94(1):110-6.
175. Sánchez A, Espinosa P, García T, Mancilla R. The 19 kDa *Mycobacterium tuberculosis* lipoprotein (LpqH) induces macrophage apoptosis through extrinsic and intrinsic pathways: a role for the mitochondrial apoptosis-inducing factor. Clin Dev Immunol. 2012;2012:950503.
176. Mohn WW, van der Geize R, Stewart GR, Okamoto S, Liu J, Dijkhuizen L, et al. The actinobacterial mce4 locus encodes a steroid transporter. J Biol Chem. 2008;283(51):35368-74.
177. Casali N, White AM, Riley LW. Regulation of the *Mycobacterium tuberculosis* mce1 operon. J Bacteriol. 2006;188(2):441-9.
178. Hunt TA, Kooi C, Sokol PA, Valvano MA. Identification of *Burkholderia cenocepacia* genes required for bacterial survival in vivo. Infect Immun. 2004;72(7):4010-22.
179. Ernst RK, D'Argenio DA, Ichikawa JK, Bangera MG, Selgrade S, Burns JL, et al. Genome mosaicism is conserved but not unique in *Pseudomonas aeruginosa* isolates from the airways of young children with cystic fibrosis : Microarray analysis of P. aeruginosa genomes. Environ Microbiol. 2003;5(12):1341-9.
180. Abdalla MY, Ahmad IM, Switzer B, Britigan BE. Induction of heme oxygenase-1 contributes to survival of *Mycobacterium abscessus* in human macrophages-like THP-1 cells. Redox Biol. 2015;4:328-39.
181. Miranda-CasoLuengo AA, Staunton PM, Dinan AM, Lohan AJ, Loftus BJ. Functional characterization of the *Mycobacterium abscessus* genome coupled with condition specific transcriptomics reveals conserved molecular strategies for host adaptation and persistence. BMC Genomics. 2016;17:553.
182. Feliziani S, Marvig RL, Luján AM, Moyano AJ, Di Rienzo JA, Krogh Johansen H, et al. Coexistence and within-host evolution of diversified lineages of hypermutable *Pseudomonas aeruginosa* in long-term cystic fibrosis infections. PLoS Genet. 2014;10(10):e1004651.
183. Kreutzfeldt KM, McAdam PR, Claxton P, Holmes A, Louise Seagar A, Laurenson IF, et al. Molecular Longitudinal Tracking of *Mycobacterium abscessus* spp. during Chronic Infection of the Human Lung. PLoS One. 2013;8(5):e63237.
184. Bange FC, Brown BA, Smaczny C, Wallace RJ, Jr., Böttger EC. Lack of transmission of *Mycobacterium abscessus* among patients with cystic fibrosis attending a single clinic. Clin Infect Dis. 2001;32(11):1648-50.
185. Harris KA, Underwood A, Kenna DTD, Brooks A, Kavaliunaite E, Kapatai G, et al. Whole-Genome Sequencing and Epidemiological Analysis Do Not Provide Evidence for Cross-transmission of *Mycobacterium abscessus* in a Cohort of Pediatric Cystic Fibrosis Patients. Clin Infect Dis. 2015;60(7):1007-16.
186. Correia AM, Ferreira JS, Borges V, Nunes A, Gomes B, Capucho R, et al. Probable Person-to-Person Transmission of Legionnaires' Disease. N Engl J Med. 2016;374(5):497-8.

187. David S, Rusniok C, Mentasti M, Gomez-Valero L, Harris SR, Lechat P, et al. Multiple major disease-associated clones of *Legionella pneumophila* have emerged recently and independently. Genome Res. 2016.
188. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953;171(4356):737-8.
189. Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, et al. Structure of a ribonucleic acid. Science. 1965;147(3664):1462-5.
190. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977;74(2):560-4.
191. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-7.
192. Kasper TJ, Melera M, Gozel P, Brownlee RG. Separation and detection of DNA by capillary electrophoresis. J Chromatogr. 1988;458:303-12.
193. Nyrén P, Lundin A. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. Anal Biochem. 1985;151(2):504-9.
194. Hyman ED. A new method of sequencing DNA. Anal Biochem. 1988;174(2):423-36.
195. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.
196. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452(7189):872-6.
197. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.
198. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13(5):278-89.
199. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009;4(4):265-70.
200. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol. 2015;33(3):296-300.
201. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228-32.
202. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre ABR, et al. Nanopore DNA Sequencing and Genome Assembly on the International Space Station. Sci Rep. 2017;7(1):18022.
203. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995;269(5223):496-512.
204. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of *Mycoplasma genitalium*. Science. 1995;270(5235):397-403.
205. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, et al. Massive gene decay in the leprosy bacillus. Nature. 2001;409(6823):35059006.
206. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Commun. 2014;5:5471.
207. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 2005;102(39):13950-5.

208. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327(5964):469-74.
209. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill F-X, Goodhead I, et al. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nature Genetics. 2008;40:987.
210. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 2011;477(7365):462-5.
211. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci U S A. 2013;110(29):11923-7.
212. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. PLoS Genet. 2014;10(8):e1004547.
213. Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun. 2016;7:12797.
214. Phelan J, de Sessions PF, Tientcheu L, Perdigao J, Machado D, Hasan R, et al. Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. Sci Rep. 2018;8(1):160.
215. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bioGN]. 2013.
216. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
217. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016;2(4):e000056.
218. Farris JW. Methods for Computing Wagner Trees. Syst Zool. 1970;19(1).
219. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821-9.
220. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
221. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986;3(5):418-26.
222. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scand Stat Theory Appl. 1979;6(2):65-70.
223. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3.
224. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):D222-30.
225. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236-40.
226. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol Biol Evol. 2017;34(8):2115-22.
227. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res. 2011;39(Web Server issue):W347-52.
228. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016;44(W1):W16-21.
229. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 2010;38(Database issue):D234-6.

230. McGuire AM, Weiner B, Park ST, Wapinski I, Raman S, Dolganov G, et al. Comparative analysis of Mycobacterium and related Actinomycetes yields insight into the evolution of *Mycobacterium tuberculosis* pathogenesis. BMC Genomics. 2012;13:120.
231. Mycobase. M. a. *abscessus* ATCC19977 GO-terms [Cited 25 September 2018]. Available from: http://strong.ucdenver.edu/mycobase.
232. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006;22(13):1600-7.
233. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008;2008:619832.
234. Sali A, Blundell TL. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J Mol Biol. 1990;212(2):403-28.
235. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol. 2001;310(1):243-57.
236. Medjahed H, Singh AK. Genetic manipulation of *Mycobacterium abscessus*. Curr Protoc Microbiol. 2010;Chapter 10:Unit 10D.2.
237. Kumar R, Mukhopadhyay AK, Rao DN. Characterization of an N6 adenine methyltransferase from *Helicobacter pylori* strain 26695 which methylates adjacent adenines on the same strand: N6 adenine methyltransferase from H. pylori 26695. FEBS J. 2010;277(7):1666-83.
238. Punekar AS, Liljeruhm J, Shepherd TR, Forster AC, Selmer M. Structural and functional insights into the molecular mechanism of rRNA m6A methyltransferase RlmJ. Nucleic Acids Res. 2013;41(20):9537-48.
239. Pacific Biosciences, SMRT Analysis algorithms for epigenitic analysis. [Cited September 25 2018]. Available from: https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/analysis-applications/epigenetics/
240. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563-9.
241. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13:238.
242. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010;7(6):461-5.
243. Cumming BM, Steyn AJC. Metabolic plasticity of central carbon metabolism protects mycobacteria. Proc Natl Acad Sci U S A. 2015;112(43):13135-6.
244. Venugopal A, Bryk R, Shi S, Rhee K, Rath P, Schnappinger D, et al. Virulence of *Mycobacterium tuberculosis* depends on lipoamide dehydrogenase, a member of three multienzyme complexes. Cell Host Microbe. 2011;9(1):21-31.
245. Basavanna S, Khandavilli S, Yuste J, Cohen JM, Hosie AHF, Webb AJ, et al. Screening of *Streptococcus pneumoniae* ABC transporter mutants demonstrates that LivJHMGF, a branched-chain amino acid ABC transporter, is necessary for disease pathogenesis. Infect Immun. 2009;77(8):3412-23.
246. Kaiser JC, Sen S, Sinha A, Wilkinson BJ, Heinrichs DE. The role of two branched-chain amino acid transporters in *Staphylococcus aureus* growth, membrane fatty acid composition and virulence. Mol Microbiol. 2016;102(5):850-64.
247. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol. 2003;48(1):77-84.
248. Bailo R, Bhatt A, Aínsa JA. Lipid transport in *Mycobacterium tuberculosis* and its implications in virulence and drug development. Biochem Pharmacol. 2015;96(3):159-67.

249. Kurtz S, McKinnon KP, Runge MS, Ting JPY, Braunstein M. The SecA2 secretion factor of *Mycobacterium tuberculosis* promotes growth in macrophages and inhibits the host immune response. Infect Immun. 2006;74(12):6855-64.
250. Lee W, VanderVen BC, Fahey RJ, Russell DG. Intracellular *Mycobacterium tuberculosis* exploits host-derived fatty acids to limit metabolic stress. J Biol Chem. 2013;288(10):6788-800.
251. Kang Y, Zarzycki-Siek J, Walton CB, Norris MH, Hoang TT. Multiple FadD acyl-CoA synthetases contribute to differential fatty acid degradation and virulence in *Pseudomonas aeruginosa*. PLoS One. 2010;5(10):e13557.
252. Hett EC, Rubin EJ. Bacterial growth and cell division: a mycobacterial perspective. Microbiol Mol Biol Rev. 2008;72(1):126-56, table of contents.
253. Mosquera-Rendón J, Rada-Bravo AM, Cárdenas-Brito S, Corredor M, Restrepo-Pineda E, Benítez-Páez A. Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species. BMC Genomics. 2016;17:45.
254. Brynildsrud OB, Eldholm V, Bohlin J, Uadiale K, Obaro S, Caugant DA. Acquisition of virulence genes by a carrier strain gave rise to the ongoing epidemics of meningococcal disease in West Africa. Proc Natl Acad Sci U S A. 2018;115(21):5510-5.
255. Senaratne RH, De Silva AD, Williams SJ, Mougous JD, Reader JR, Zhang T, et al. 5′-Adenosinephosphosulphate reductase (CysH) protects *Mycobacterium tuberculosis* against free radicals during chronic infection phase in mice. Mol Microbiol. 2006;59(6):1744-53.
256. Shin D-M, Jeon B-Y, Lee H-M, Jin HS, Yuk J-M, Song C-H, et al. *Mycobacterium tuberculosis* eis regulates autophagy, inflammation, and cell death through redox-dependent signaling. PLoS Pathog. 2010;6(12):e1001230.
257. Garzan A, Willby MJ, Ngo HX, Gajadeera CS, Green KD, Holbrook SYL, et al. Combating Enhanced Intracellular Survival (Eis)-Mediated Kanamycin Resistance of *Mycobacterium tuberculosis* by Novel Pyrrolo[1,5-a]pyrazine-Based Eis Inhibitors. ACS Infect Dis. 2017;3(4):302-9.
258. Aussel L, Pierrel F, Loiseau L, Lombard M, Fontecave M, Barras F. Biosynthesis and physiology of coenzyme Q in bacteria. Biochim Biophys Acta. 2014;1837(7):1004-11.
259. Casali N, Riley LW. A phylogenomic analysis of the Actinomycetales mce operons. BMC Genomics. 2007;8:60.
260. Perkowski EF, Miller BK, McCann JR, Sullivan JT, Malik S, Allen IC, et al. An orphaned Mce-associated membrane protein of *Mycobacterium tuberculosis* is a virulence factor that stabilizes Mce transporters. Mol Microbiol. 2016;100(1):90-107.
261. Eichhorn E, van der Ploeg JR, Leisinger T. Deletion analysis of the *Escherichia coli* taurine and alkanesulfonate transport systems. J Bacteriol. 2000;182(10):2687-95.
262. Chauhan R, Mande SC. Site-directed mutagenesis reveals a novel catalytic mechanism of *Mycobacterium tuberculosis* alkylhydroperoxidase C. Biochem J. 2002;367(Pt 1):255-61.
263. Wong CF, Shin J, Subramanian Manimekalai MS, Saw WG, Yin Z, Bhushan S, et al. AhpC of the mycobacterial antioxidant defense system and its interaction with its reducing partner Thioredoxin-C. Sci Rep. 2017;7(1):5159.
264. Sachdeva P, Misra R, Tyagi AK, Singh Y. The sigma factors of *Mycobacterium tuberculosis*: regulation of the regulators: The σ-factors of *M. tuberculosis*. FEBS J. 2010;277(3):605-26.
265. Howard ST, Newman KL, McNulty S, Brown-Elliott BA, Vasireddy R, Bridge L, et al. Insertion site and distribution of a genomic island conferring DNA phosphorothioation in the *Mycobacterium abscessus* complex. Microbiology. 2013;159(Pt 11):2323-32.
266. Xie X, Liang J, Pu T, Xu F, Yao F, Yang Y, et al. Phosphorothioate DNA as an antioxidant in bacteria. Nucleic Acids Res. 2012;40(18):9115-24.
267. Law RJ, Hamlin JNR, Sivro A, McCorrister SJ, Cardama GA, Cardona ST. A functional phenylacetic acid catabolic pathway is required for full pathogenicity of *Burkholderia cenocepacia* in the *Caenorhabditis elegans* host model. J Bacteriol. 2008;190(21):7209-18.

268. Tischler AD, Leistikow RL, Kirksey MA, Voskuil MI, McKinney JD. *Mycobacterium tuberculosis* requires phosphate-responsive gene regulation to resist host immunity. Infect Immun. 2013;81(1):317-28.
269. Brokaw AM, Eide BJ, Muradian M, Boster JM, Tischler AD. *Mycobacterium smegmatis* PhoU Proteins Have Overlapping Functions in Phosphate Signaling and Are Essential. Front Microbiol. 2017;8:2523.
270. Stahl C, Kubetzko S, Kaps I, Seeber S, Engelhardt H, Niederweis M. MspA provides the main hydrophilic pathway through the cell wall of *Mycobacterium smegmatis*. Mol Microbiol. 2001;40(2):451-64.
271. Blanco P, Hernando-Amado S, Reales-Calderon JA, Corona F, Lira F, Alcalde-Rico M, et al. Bacterial Multidrug Efflux Pumps: Much More Than Antibiotic Resistance Determinants. Microorganisms. 2016;4(1).
272. Gioffré A, Infante E, Aguilar D, Santangelo MDlP, Klepp L, Amadio A, et al. Mutation in mce operons attenuates *Mycobacterium tuberculosis* virulence. Microbes Infect. 2005;7(3):325-34.
273. Kumar A, Bose M, Brahmachari V. Analysis of Expression Profile of Mammalian Cell Entry (*mce*) Operons of *Mycobacterium tuberculosis*. Infection and Immunity. 2003;71(10):6083.
274. Forde BM, Phan M-D, Gawthorne JA, Ashcroft MM, Stanton-Cook M, Sarkar S, et al. Lineage-Specific Methyltransferases Define the Methylome of the Globally Disseminated *Escherichia coli* ST131 Clone. MBio. 2015;6(6):e01602-15.
275. Torrents E. Ribonucleotide reductases: essential enzymes for bacterial life. Front Cell Infect Microbiol. 2014;4:52.
276. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27(7):1009-10.
277. Bheemanaik S, Reddy YVR, Rao DN. Structure, function and mechanism of exocyclic DNA methyltransferases. Biochem J. 2006;399(2):177-90.
278. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol. 2008;6(12):e311.
279. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 2006;239(2):226-35.
280. Castillo-Ramírez S, Harris SR, Holden MTG, He M, Parkhill J, Bentley SD, et al. The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog. 2011;7(7):e1002129.
281. He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, et al. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. Nat Genet. 2013;45(1):109-13.
282. Aussel L, Loiseau L, Hajj Chehade M, Pocachard B, Fontecave M, Pierrel F, et al. ubiJ, a new gene required for aerobic growth and proliferation in macrophage, is involved in coenzyme Q biosynthesis in *Escherichia coli* and Salmonella enterica serovar Typhimurium. J Bacteriol. 2014;196(1):70-9.
283. Agrawal S, Jaswal K, Shiver AL, Balecha H, Patra T, Chaba R. A genome-wide screen in *Escherichia coli* reveals that ubiquinone is a key antioxidant for metabolism of long-chain fatty acids. J Biol Chem. 2017;292(49):20086-99.
284. Schnappinger D, Ehrt S, Voskuil MI, Liu Y, Mangan JA, Monahan IM, et al. Transcriptional Adaptation of *Mycobacterium tuberculosis* within Macrophages: Insights into the Phagosomal Environment. J Exp Med. 2003;198(5):693-704.
285. Niederweis M. Mycobacterial porins--new channel proteins in unique outer membranes. Mol Microbiol. 2003;49(5):1167-77.
286. Sharbati-Tehrani S. Porins limit the intracellular persistence of *Mycobacterium smegmatis*. Microbiology. 2005;151(7):2403-10.

287. Svetlíková Z, Skovierová H, Niederweis M, Gaillard J-L, McDonnell G, Jackson M. Role of porins in the susceptibility of *Mycobacterium smegmatis* and *Mycobacterium chelonae* to aldehyde-based disinfectants and drugs. Antimicrob Agents Chemother. 2009;53(9):4015-8.
288. Gesbert G, Ramond E, Tros F, Dairou J, Frapy E, Barel M, et al. Importance of branched-chain amino acid utilization in Francisella intracellular adaptation. Infect Immun. 2015;83(1):173-83.
289. Hondalus MK, Bardarov S, Russell R, Chan J, Jacobs WR, Jr., Bloom BR. Attenuation of and protection induced by a leucine auxotroph of *Mycobacterium tuberculosis*. Infect Immun. 2000;68(5):2888-98.
290. Pandey AK, Sassetti CM. Mycobacterial persistence requires the utilization of host cholesterol. Proc Natl Acad Sci U S A. 2008;105(11):4376-80.
291. Miller WG, Pearson BM, Wells JM, Parker CT, Kapitonov VV, Mandrell RE. Diversity within the *Campylobacter jejuni* type I restriction–modification loci. Microbiology. 2005;151(2):337-51.
292. Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. 2013;41(8):4360-77.
293. Vasu K, Nagaraja V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. Microbiol Mol Biol Rev. 2013;77(1):53-72.
294. Pacific biosciences. Base Modification : From Sequencing Data to a High Confidence Motif List 2014 [Cited September 25 2018]. Available from: https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Base-Modification-:--From-Sequencing-Data-to-a-High-Confidence-Motif-List.
295. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455(7216):1069-75.
296. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279-85.
297. Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J. ACT: the Artemis comparison tool. Bioinformatics. 2005;21(16):3422-3.
298. Ishikawa J, Yamashita A, Mikami Y, Hoshino Y, Kurita H, Hotta K, et al. The complete genomic sequence of Nocardia farcinica IFM 10152. Proc Natl Acad Sci U S A. 2004;101(41):14925-30.
299. Arruda S, Bomfim G, Knights R, Huima-Byron T, Riley LW. Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells. Science. 1993;261(5127):1454-7.
300. Forrellad MA, Bianco MV, Blanco FC, Nuñez J, Klepp LI, Vazquez CL, et al. Study of the in vivo role of Mce2R, the transcriptional regulator of mce2 operon in *Mycobacterium tuberculosis*. BMC Microbiol. 2013;13:200.
301. Klepp LI, Forrellad MA, Osella AV, Blanco FC, Stella EJ, Bianco MV, et al. Impact of the deletion of the six mce operons in *Mycobacterium smegmatis*. Microbes Infect. 2012;14(7-8):590-9.
302. Clark LC, Seipke RF, Prieto P, Willemse J, van Wezel GP, Hutchings MI, et al. Mammalian cell entry genes in Streptomyces may provide clues to the evolution of bacterial virulence. Sci Rep. 2013;3:1109.
303. de la Paz Santangelo M, Klepp L, Nuñez-García J, Blanco FC, Soria M, García-Pelayo MdC, et al. Mce3R, a TetR-type transcriptional repressor, controls the expression of a regulon involved in lipid metabolism in *Mycobacterium tuberculosis*. Microbiology. 2009;155(Pt 7):2245-55.
304. Uchida Y, Casali N, White A, Morici L, Kendall LV, Riley LW. Accelerated immunopathological response of mice infected with *Mycobacterium tuberculosis* disrupted in the mce1 operon negative transcriptional regulator. Cell Microbiol. 2007;9(5):1275-83.

305. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. Nat Genet. 2015;47(3):242-9.
306. van Ingen J, Boeree MJ, van Soolingen D, Mouton JW. Resistance mechanisms and drug susceptibility testing of nontuberculous mycobacteria. Drug Resist Updat. 2012;15(3):149-61.
307. Marvig RL, Damkiær S, Khademi SMH, Markussen TM, Molin S, Jelsbak L. Within-host evolution of *Pseudomonas aeruginosa* reveals adaptation toward iron acquisition from hemoglobin. MBio. 2014;5(3):e00966-14.
308. Marvig RL, Sommer LM, Molin S, Johansen HK. Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. Nat Genet. 2015;47(1):57-64.
309. Maliniak ML, Stecenko AA, McCarty NA. A longitudinal analysis of chronic MRSA and *Pseudomonas aeruginosa* co-infection in cystic fibrosis: A single-center study. J Cyst Fibros. 2016;15(3):350-6.
310. McAdam PR, Holmes A, Templeton KE, Fitzgerald JR. Adaptive evolution of *Staphylococcus aureus* during chronic endobronchial infection of a cystic fibrosis patient. PLoS One. 2011;6(9):e24301.
311. Lee AH-Y, Flibotte S, Sinha S, Paiero A, Ehrlich RL, Balashov S, et al. Phenotypic diversity and genotypic flexibility of *Burkholderia cenocepacia* during long-term chronic infection of cystic fibrosis lungs. Genome Res. 2017;27(4):650-62.
312. Silva IN, Santos PM, Santos MR, Zlosnik JEA, Speert DP, Buskirk SW, et al. Long-Term Evolution of Burkholderia multivorans during a Chronic Cystic Fibrosis Infection Reveals Shifting Forces of Selection. mSystems. 2016;1(3):e00029-16.
313. Maurer FP, Rüegger V, Ritter C, Bloemberg GV, Böttger EC. Acquisition of clarithromycin resistance mutations in the 23S rRNA gene of *Mycobacterium abscessus* in the presence of inducible erm(41). J Antimicrob Chemother. 2012;67(11):2606-11.
314. Broset E, Martín C, Gonzalo-Asensio J. Evolutionary landscape of the *Mycobacterium tuberculosis* complex from the viewpoint of PhoPR: implications for virulence regulation and application to vaccine development. MBio. 2015;6(5):e01289-15.
315. Gonzalo-Asensio J, Mostowy S, Harders-Westerveen J, Huygen K, Hernández-Pando R, Thole J, et al. PhoP: a missing piece in the intricate puzzle of *Mycobacterium tuberculosis* virulence. PLoS One. 2008;3(10):e3496.
316. van Kessel JC, Hatfull GF. Recombineering in *Mycobacterium tuberculosis*. Nat Methods. 2007;4(2):147-52.
317. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
318. Picard Tools. [Cited September 25 2018] Available from: http://broadinstitute.github.io/picard/
319. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-8.
320. Wellcome Trust Sanger Institute Pathogen Informatics. Bio-RNASeq [Cited September 25 2018]. Available from: https://github.com/sanger-pathogens/Bio-RNASeq
321. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
322. Wellcome Sanger Institue Pathogen Informatics. Deago 2018 [Cited September 25 2108]. Available from: https://github.com/sanger-pathogens/deago#introduction.
323. Cunningham RP, Weiss B. Endonuclease III (nth) mutants of *Escherichia coli*. Proc Natl Acad Sci U S A. 1985;82(2):474-8.
324. Saito Y, Uraki F, Nakajima S, Asaeda A, Ono K, Kubo K, et al. Characterization of endonuclease III (nth) and endonuclease VIII (nei) mutants of *Escherichia coli* K-12. J Bacteriol. 1997;179(11):3783-5.

325. Chou CJ, Wisedchaisri G, Monfeli RR, Oram DM, Holmes RK, Hol WGJ, et al. Functional studies of the *Mycobacterium tuberculosis* iron-dependent regulator. J Biol Chem. 2004;279(51):53554-61.
326. Love JF, vanderSpek JC, Marin V, Guerrero L, Logan TM, Murphy JR. Genetic and biophysical studies of diphtheria toxin repressor (DtxR) and the hyperactive mutant DtxR(E175K) support a multistep model of activation. Proc Natl Acad Sci U S A. 2004;101(8):2506-11.
327. Manabe YC, Hatem CL, Kesavan AK, Durack J, Murphy JR. Both Corynebacterium diphtheriae DtxR(E175K) and *Mycobacterium tuberculosis* IdeR(D177K) are dominant positive repressors of IdeR-regulated genes in *M. tuberculosis*. Infect Immun. 2005;73(9):5988-94.
328. Ranganathan S, Cheung J, Cassidy M, Ginter C, Pata JD, McDonough KA. Novel structural features drive DNA binding properties of Cmr, a CRP family protein in TB complex mycobacteria. Nucleic Acids Res. 2018;46(1):403-20.
329. Kudhair BK, Hounslow AM, Rolfe MD, Crack JC, Hunt DM, Buxton RS, et al. Structure of a Wbl protein and implications for NO sensing by *M. tuberculosis*. Nat Commun. 2017;8(1):2280.
330. Smith LJ, Stapleton MR, Buxton RS, Green J. Structure-function relationships of the *Mycobacterium tuberculosis* transcription factor WhiB1. PLoS One. 2012;7(7):e40407.
331. Baker JJ, Johnson BK, Abramovitch RB. Slow growth of *Mycobacterium tuberculosis* at acidic pH is regulated by phoPR and host-associated carbon sources. Mol Microbiol. 2014;94(1):56-69.
332. Sanz J, Navarro J, Arbués A, Martín C, Marijuán PC, Moreno Y. The transcriptional regulatory network of *Mycobacterium tuberculosis*. PLoS One. 2011;6(7):e22178.
333. Gerasimova A, Kazakov AE, Arkin AP, Dubchak I, Gelfand MS. Comparative genomics of the dormancy regulons in mycobacteria. J Bacteriol. 2011;193(14):3446-52.
334. Hatzios SK, Bertozzi CR. The regulation of sulfur metabolism in *Mycobacterium tuberculosis*. PLoS Pathog. 2011;7(7):e1002036.
335. Grinberg I, Shteinberg T, Gorovitz B, Aharonowitz Y, Cohen G, Borovok I. The Streptomyces NrdR transcriptional regulator is a Zn ribbon/ATP cone protein that binds to the promoter regions of class Ia and class II ribonucleotide reductase operons. J Bacteriol. 2006;188(21):7635-44.
336. Mowa MB, Warner DF, Kaplan G, Kana BD, Mizrahi V. Function and regulation of class I ribonucleotide reductase-encoding genes in mycobacteria. J Bacteriol. 2009;191(3):985-95.
337. Anil Kumar V, Goyal R, Bansal R, Singh N, Sevalkar RR, Kumar A, et al. EspR-dependent ESAT-6 Protein Secretion of *Mycobacterium tuberculosis* Requires the Presence of Virulence Regulator PhoP. J Biol Chem. 2016;291(36):19018-30.
338. Mishra AK, Driessen NN, Appelmelk BJ, Besra GS. Lipoarabinomannan and related glycoconjugates: structure, biogenesis and role in *Mycobacterium tuberculosis* physiology and host-pathogen interaction. FEMS Microbiol Rev. 2011;35(6):1126-57.
339. Kim H-Y, Kim BJ, Kook Y, Yun Y-J, Shin JH, Kim B-J, et al. Mycobacterium *massiliense* is differentiated from *Mycobacterium abscessus* and Mycobacterium *bolletii* by erythromycin ribosome methyltransferase gene (erm) and clarithromycin susceptibility patterns. Microbiol Immunol. 2010;54(6):347-53.
340. Rubio M, March F, Garrigó M, Moreno C, Español M, Coll P. Inducible and Acquired Clarithromycin Resistance in the *Mycobacterium abscessus* Complex. PLoS One. 2015;10(10):e0140166.
341. Meier A, Kirschner P, Springer B, Steingrube VA, Brown BA, Wallace RJ, Jr., et al. Identification of mutations in 23S rRNA gene of clarithromycin-resistant Mycobacterium intracellulare. Antimicrob Agents Chemother. 1994;38(2):381-4.
342. Pricer RE, Houghton JL, Green KD, Mayhoub AS, Garneau-Tsodikova S. Biochemical and structural analysis of aminoglycoside acetyltransferase Eis from *Anabaena variabilis*. Mol Biosyst. 2012;8(12):3305-13.

343. Prammananan T, Sander P, Brown BA, Frischkorn K, Onyi GO, Zhang Y, et al. A single 16S ribosomal RNA substitution is responsible for resistance to amikacin and other 2-deoxystreptamine aminoglycosides in *Mycobacterium abscessus* and *Mycobacterium chelonae*. J Infect Dis. 1998;177(6):1573-81.
344. Mena A, Smith EE, Burns JL, Speert DP, Moskowitz SM, Perez JL, et al. Genetic adaptation of *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients is catalyzed by hypermutation. J Bacteriol. 2008;190(24):7910-7.
345. Ferroni A, Guillemot D, Moumile K, Bernede C, Le Bourgeois M, Waernessyckle S, et al. Effect of mutator P. aeruginosa on antibiotic resistance acquisition and respiratory function in cystic fibrosis. Pediatr Pulmonol. 2009;44(8):820-5.
346. Hogardt M, Hoboth C, Schmoldt S, Henke C, Bader L, Heesemann J. Stage-specific adaptation of hypermutable *Pseudomonas aeruginosa* isolates during chronic pulmonary infection in patients with cystic fibrosis. J Infect Dis. 2007;195(1):70-80.
347. Montanari S, Oliver A, Salerno P, Mena A, Bertoni G, Tümmler B, et al. Biological cost of hypermutation in *Pseudomonas aeruginosa* strains from patients with cystic fibrosis. Microbiology. 2007;153(Pt 5):1445-54.
348. Wanner RM, Castor D, Güthlein C, Böttger EC, Springer B, Jiricny J. The uracil DNA glycosylase UdgB of *Mycobacterium smegmatis* protects the organism from the mutagenic effects of cytosine and adenine deamination. J Bacteriol. 2009;191(20):6312-9.
349. Manganelli R, Proveddi R, Rodrigue S, Beaucher J, Gaudreau L, Smith I. σ Factors and Global Gene Regulation in *Mycobacterium tuberculosis*. Journal of Bacteriology. 2004;186(4):895.
350. Rodriguez GM, Voskuil MI, Gold B, Schoolnik GK, Smith I. ideR, An essential gene in *Mycobacterium tuberculosis*: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. Infect Immun. 2002;70(7):3371-81.
351. Bai G, Knapp GS, McDonough KA. Cyclic AMP signalling in mycobacteria: redirecting the conversation with a common currency. Cell Microbiol. 2011;13(3):349-58.
352. Welin A, Winberg ME, Abdalla H, Särndahl E, Rasmusson B, Stendahl O, et al. Incorporation of *Mycobacterium tuberculosis* lipoarabinomannan into macrophage membrane rafts is a prerequisite for the phagosomal maturation block. Infect Immun. 2008;76(7):2882-7.
353. Bai G, McCue LA, McDonough KA. Characterization of *Mycobacterium tuberculosis* Rv3676 (CRPMt), a cyclic AMP receptor protein-like DNA binding protein. J Bacteriol. 2005;187(22):7795-804.
354. Rohde KH, Abramovitch RB, Russell DG. *Mycobacterium tuberculosis* invasion of macrophages: linking bacterial gene expression to environmental cues. Cell Host Microbe. 2007;2(5):352-64.
355. Abou Alaiwa MH, Beer AM, Pezzulo AA, Launspach JL, Horan RA, Stoltz DA, et al. Neonates with Cystic Fibrosis Have a Reduced Nasal Liquid pH; A Small Pilot Study. Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society. 2014;13(4):373-7.
356. Zondervan NA, van Dam JCJ, Schaap PJ, Martins Dos Santos VAP, Suarez-Diez M. Regulation of Three Virulence Strategies of *Mycobacterium tuberculosis*: A Success Story. Int J Mol Sci. 2018;19(2).
357. Rickman L, Scott C, Hunt DM, Hutchinson T, Menéndez MC, Whalan R, et al. A member of the cAMP receptor protein family of transcription regulators in *Mycobacterium tuberculosis* is required for virulence in mice and controls transcription of the rpfA gene coding for a resuscitation promoting factor. Molecular Microbiology. 2005;56(5):1274-86.
358. Reddy MCM, Palaninathan SK, Bruning JB, Thurman C, Smith D, Sacchettini JC. Structural insights into the mechanism of the allosteric transitions of *Mycobacterium tuberculosis* cAMP receptor protein. J Biol Chem. 2009;284(52):36581-91.
359. Frigui W, Bottai D, Majlessi L, Monot M, Josselin E, Brodin P, et al. Control of *M. tuberculosis* ESAT-6 secretion and specific T cell recognition by PhoP. PLoS Pathog. 2008;4(2):e33.

360. Walters SB, Dubnau E, Kolesnikova I, Laval F, Daffe M, Smith I. The *Mycobacterium tuberculosis* PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis. Mol Microbiol. 2006;60(2):312-30.
361. Gonzalo-Asensio J, Malaga W, Pawlik A, Astarie-Dequeker C, Passemar C, Moreau F, et al. Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. Proc Natl Acad Sci U S A. 2014;111(31):11491-6.
362. Glover RT, Kriakov J, Garforth SJ, Baughn AD, Jacobs WR, Jr. The two-component regulatory system senX3-regX3 regulates phosphate-dependent gene expression in *Mycobacterium smegmatis*. J Bacteriol. 2007;189(15):5495-503.
363. Mikonranta L, Mappes J, Laakso J, Ketola T. Within-host evolution decreases virulence in an opportunistic bacterial pathogen. BMC Evol Biol. 2015;15:165.
364. Smith EE, Buckley DG, Wu Z, Saenphimmachak C, Hoffman LR, D'Argenio DA, et al. Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. Proc Natl Acad Sci U S A. 2006;103(22):8487-92.
365. Blasco B, Chen JM, Hartkoorn R, Sala C, Uplekar S, Rougemont J, et al. Virulence regulator EspR of *Mycobacterium tuberculosis* is a nucleoid-associated protein. PLoS Pathog. 2012;8(3):e1002621.
366. Panosa A, Roca I, Gibert I. Ribonucleotide reductases of *Salmonella typhimurium*: transcriptional regulation and differential role in pathogenesis. PLoS One. 2010;5(6):e11328.
367. Naveen V, Hsiao C-D. NrdR Transcription Regulation: Global Proteome Analysis and Its Role in *Escherichia coli* Viability and Virulence. PLoS One. 2016;11(6):e0157165.
368. Salzman V, Mondino S, Sala C, Cole ST, Gago G, Gramajo H. Transcriptional regulation of lipid homeostasis in mycobacteria. Mol Microbiol. 2010;78(1):64-77.
369. Wei J, Dahl JL, Moulder JW, Roberts EA, O'Gaora P, Young DB, et al. Identification of a *Mycobacterium tuberculosis* gene that enhances mycobacterial survival in macrophages. J Bacteriol. 2000;182(2):377-84.
370. Zaunbrecher MA, Sikes RD, Jr., Metchock B, Shinnick TM, Posey JE. Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A. 2009;106(47):20004-9.
371. He L, Wang X, Cui P, Jin J, Chen J, Zhang W, et al. *ubiA* (Rv3806c) encoding DPPR synthase involved in cell wall synthesis is associated with ethambutol resistance in *Mycobacterium tuberculosis*. Tuberculosis. 2015;95(2):149-54.
372. Wang L, Xu M, Southall N, Zheng W, Wang S. A High-Throughput Assay for Developing Inhibitors of PhoP, a Virulence Factor of *Mycobacterium tuberculosis*. Combinatorial chemistry & high throughput screening. 2016;19(10):855-64.
373. Duarte RS, Lourenço MCS, Fonseca LdS, Leão SC, Amorim EdLT, Rocha ILL, et al. Epidemic of postsurgical infections caused by Mycobacterium *massiliense*. J Clin Microbiol. 2009;47(7):2149-55.
374. Viana-Niero C, Lima KVB, Lopes ML, da Silva Rabello MC, Marsola LR, Brilhante VCR, et al. Molecular Characterization of Mycobacterium *massiliense* and Mycobacterium *bolletii* in Isolates Collected from Outbreaks of Infections after Laparoscopic Surgeries and Cosmetic Procedures. J Clin Microbiol. 2008;46(3):850-5.
375. Nunes LdS, Baethgen LF, Ribeiro MO, Cardoso CM, de Paris F, De David SMM, et al. Outbreaks due to *Mycobacterium abscessus* subsp. *bolletii* in southern Brazil: persistence of a single clone from 2007 to 2011. J Med Microbiol. 2014;63(Pt 10):1288-93.
376. Baruque Villar G, de Mello Freitas FT, Pais Ramos J, Dias Campos CE, de Souza Caldas PC, Santos Bordalo F, et al. Risk Factors for *Mycobacterium abscessus* subsp. *bolletii* infection after laparoscopic surgery during an outbreak in Brazil. Infect Control Hosp Epidemiol. 2015;36(1):81-6.
377. Cardoso AM, Martins de Sousa E, Viana-Niero C, Bonfim de Bortoli F, Pereira das Neves ZC, Leão SC, et al. Emergence of nosocomial Mycobacterium *massiliense* infection in Goiás, Brazil. Microbes Infect. 2008;10(14-15):1552-7.

378. Monego F, Duarte RS, Nakatani SM, Araújo WN, Riediger IN, Brockelt S, et al. Molecular identification and typing of Mycobacterium *massiliense* isolated from postsurgical infections in Brazil. Braz J Infect Dis. 2011;15(5):436-41.
379. Shang S, Gibbs S, Henao-Tamayo M, Shanley CA, McDonnell G, Duarte RS, et al. Increased virulence of an epidemic strain of Mycobacterium *massiliense* in mice. PLoS One. 2011;6(9):e24726.
380. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24(11):1403-5.
381. Kamvar ZN, Tabima JF, Grünwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ. 2014;2:e281.
382. Murray GGR, Wang F, Harrison EM, Paterson GK, Mather AE, Harris SR, et al. The effect of genetic structure on molecular dating and tests for temporal signal. Methods Ecol Evol. 2016;7(1):80-9.
383. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43(3):e15.
384. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7(1):214.
385. Tracer | BEAST Documentation.
386. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.
387. Ummels R, Abdallah AM, Kuiper V, Aâjoud A, Sparrius M, Naeem R, et al. Identification of a novel conjugative plasmid in mycobacteria that requires both type IV and type VII secretion. MBio. 2014;5(5):e01744-14.
388. Dumas E, Christina Boritsch E, Vandenbogaert M, Rodríguez de la Vega RC, Thiberge J-M, Caro V, et al. Mycobacterial Pan-Genome Analysis Suggests Important Role of Plasmids in the Radiation of Type VII Secretion Systems. Genome Biol Evol. 2016;8(2):387-402.
389. de Moura VCN, Gibbs S, Jackson M. Gene Replacement in *Mycobacterium chelonae*: Application to the Construction of Porin Knock-Out Mutants. PLoS One. 2014;9(4):e94951.
390. Stinear TP, Seemann T, Pidot S, Frigui W, Reysset G, Garnier T, et al. Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. Genome Res. 2007;17(2):192-200.
391. Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, et al. Intracontinental spread of human invasive *Salmonella typhimurium* pathovariants in sub-Saharan Africa. Nat Genet. 2012;44(11):ng.2423.
392. Klemm EJ, Gkrania-Klotsas E, Hadfield J, Forbester JL, Harris SR, Hale C, et al. Emergence of host-adapted Salmonella Enteritidis through rapid evolution in an immunocompromised host. Nature Microbiology. 2016;1(3):nmicrobiol201523.
393. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. Mobility of plasmids. Microbiol Mol Biol Rev. 2010;74(3):434-52.
394. Abdallah AM, Gey van Pittius NC, Champion PAD, Cox J, Luirink J, Vandenbroucke-Grauls CMJE, et al. Type VII secretion--mycobacteria show the way. Nat Rev Microbiol. 2007;5(11):883-91.
395. Serafini A, Boldrin F, Palù G, Manganelli R. Characterization of a *Mycobacterium tuberculosis* ESX-3 Conditional Mutant: Essentiality and Rescue by Iron and Zinc. J Bacteriol. 2009;191(20):6340-4.
396. Bottai D, Di Luca M, Majlessi L, Frigui W, Simeone R, Sayes F, et al. Disruption of the ESX-5 system of *Mycobacterium tuberculosis* causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. Mol Microbiol. 2012;83(6):1195-209.
397. Siegrist MS, Steigedal M, Ahmad R, Mehra A, Dragset MS, Schuster BM, et al. Mycobacterial Esx-3 requires multiple components for iron acquisition. MBio. 2014;5(3):e01073-14.

398. Majlessi L, Prados-Rosales R, Casadevall A, Brosch R. Release of mycobacterial antigens. Immunol Rev. 2015;264(1):25-45.
399. Rominski A, Selchow P, Becker K, Brülle JK, Dal Molin M, Sander P. Elucidation of *Mycobacterium abscessus* aminoglycoside and capreomycin resistance by targeted deletion of three putative resistance genes. J Antimicrob Chemother. 2017;72(8):2191-200.
400. Cheng A, Sheng WH, Huang YC, Sun HY, Tsai YT, Chen ML, et al. Prolonged postprocedural outbreak of Mycobacterium *massiliense* infections associated with ultrasound transmission gel. Clin Microbiol Infect. 2016;22(4):382.e1-.e11.
401. Simões M, Pereira MO, Machado I, Simões LC, Vieira MJ. Comparative antibacterial potential of selected aldehyde-based biocides and surfactants against planktonic *Pseudomonas fluorescens*. J Ind Microbiol Biotechnol. 2006;33(9):741-9.
402. Tschudin-Sutter S, Frei R, Kampf G, Tamm M, Pflimlin E, Battegay M, et al. Emergence of glutaraldehyde-resistant *Pseudomonas* aeruginosa. Infect Control Hosp Epidemiol. 2011;32(12):1173-8.
403. Vikram A, Bomberger JM, Bibby KJ. Efflux as a glutaraldehyde resistance mechanism in *Pseudomonas fluorescens* and *Pseudomonas* aeruginosa biofilms. Antimicrob Agents Chemother. 2015;59(6):3433-40.
404. Viljoen A, Blaise M, de Chastellier C, Kremer L. MAB_3551c encodes the primary triacylglycerol synthase involved in lipid accumulation in *Mycobacterium abscessus*. Mol Microbiol. 2016;102(4):611-27.
405. Son MS, Matthews WJ, Jr., Kang Y, Nguyen DT, Hoang TT. In vivo evidence of *Pseudomonas* aeruginosa nutrient acquisition and pathogenesis in the lungs of cystic fibrosis patients. Infect Immun. 2007;75(11):5313-24.
406. Shell SS, Prestwich EG, Baek S-H, Shah RR, Sassetti CM, Dedon PC, et al. DNA methylation impacts gene expression and ensures hypoxic survival of *Mycobacterium tuberculosis*. PLoS Pathog. 2013;9(7):e1003419.
407. Shimono N, Morici L, Casali N, Cantrell S, Sidders B, Ehrt S, et al. Hypervirulent mutant of *Mycobacterium tuberculosis* resulting from disruption of the mce1 operon. Proc Natl Acad Sci U S A. 2003;100(26):15918-23.
408. Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, Harris SR, et al. Robust high throughput prokaryote de novo assembly and improvement pipeline for Illumina data. bioRxiv. 2016.
409. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27(4):578-9.
410. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol. 2012;13(6):R56.
411. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
412. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100-8.
413. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32(1):11-6.
414. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785-6.
415. Kolbe DL, Eddy SR. Fast filtering for RNA homology search. Bioinformatics. 2011;27(22):3102-9.
416. Garcia BJ, Datta G, Davidson RM, Strong M. MycoBASE: expanding the functional annotation coverage of mycobacterial genomes. BMC Genomics. 2015;16:1102.
417. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. Bioinformatics. 2007;23(22):3024-31.